Elaborazione del Linguaggio Naturale: Interpretazione, Ragionamento automatico e Apprendimento delle macchine

Roberto Basili

(Università di Roma, Tor Vergata)

dblp: http://dblp.uni-trier.de/pers/hd/b/Basili:Roberto.html

Google scholar: https://scholar.google.com/citations?user=U1A22fYAAAAJ&hl=it&oi=sra

Università di Bologna, 16 Maggio 2016

Overview

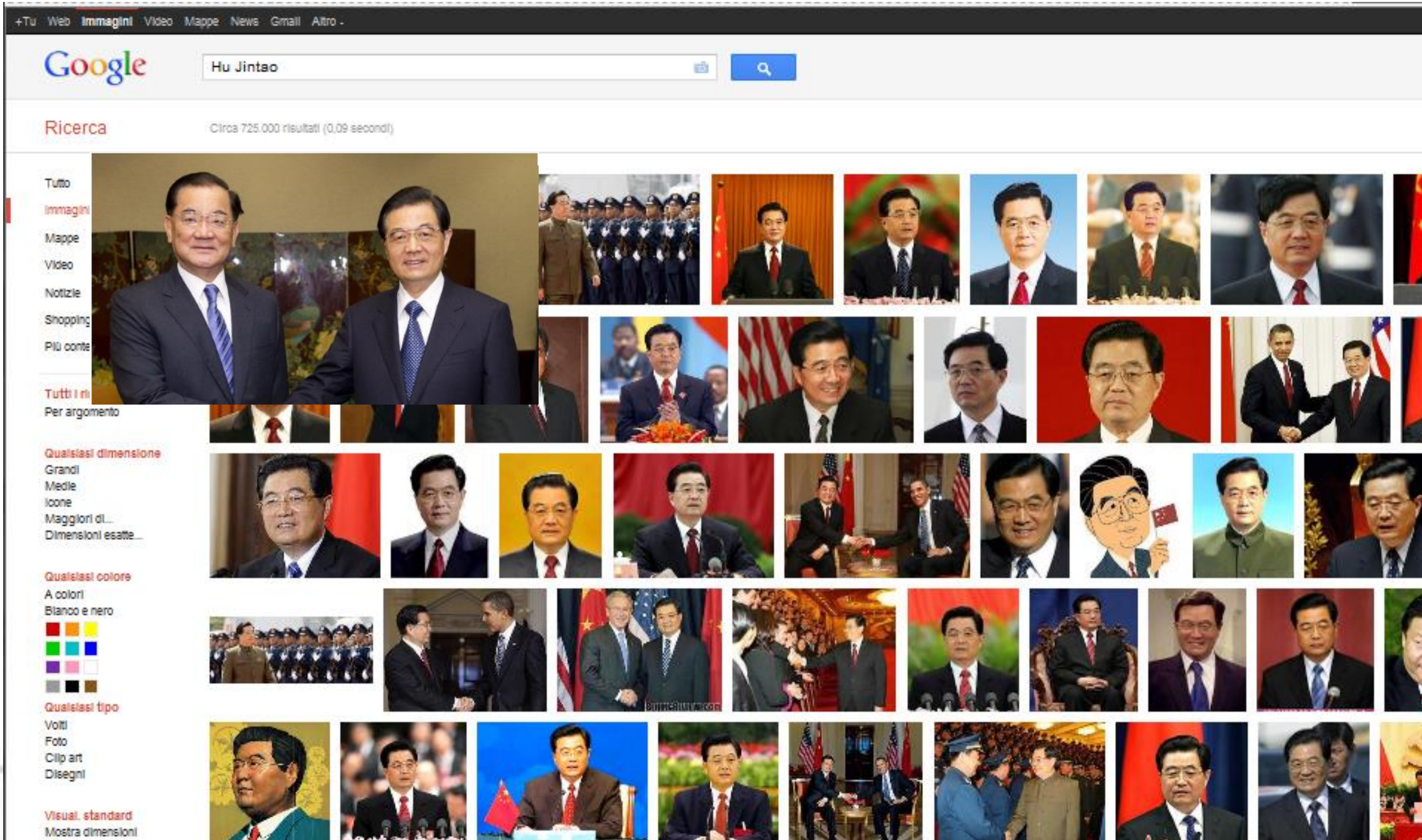
- Intelligenza Artificiale e Lingue parlate e scritte
  - Informazioni e Rappresentazioni coinvolte
  - Sfide (ri)correnti, battaglie (già) vinte e rischi inerenti …
- Elaborazione Automatica delle Lingue: Modelli, Metodi e *Risultati*
- *break*
- Ruolo delle Tecnologie dell'Apprendimento ed Applicazioni:
  - Sviluppo Automatico di Dizionari, Lessici Semantici ed Ontologie
  - Riconoscimento di fenomeni semantici
  - Trattamento Semantico della Documentazione Investigativa
  - Sistemi Web-based di Opinion Mining, Market Watch & Brand Reputation Management

Semantics, Open Data and Natural Language

Web contents, characterized by rich multimedia information, are mostly opaque from a semantic standpoint

Information, Web and Natural Languages

Who is Hu Jintao?



+Tu Web Immagini Video Mappe News Gmail Altro -
Google
Hu Jintao
Ricerca
Circa 725.000 risultati (0,09 secondi)
Tutto
Mappe
Video
Notizie
Per argomento
Qualsiasi dimensione
Grandi
Medie
Icone
Maggiori di...
Dimensioni esatte...
Qualsiasi colore
A colori
Bianco e nero
Qualsiasi tipo
Volti
Foto
Clip art
Disegni
Visual. standard
Mostra dimensioni

Content Semantics and Natural Language

- Human languages are the main carrier of the information involved in processes such as *retrieval*, *publication* and *exchange* of knowledge as it is associated to the open Web contents
- Words and NL syntactic structures express concepts, activities, events, abstractions and conceptual relations we usually share through data
- "*Language is parasitic to knowledge representation languages but the viceversa is not true*" (Wilks, 2001)
- From **Learning to Read** to **Knowledge Distillation** as a (integrated pool of) Semantic interpretation Task(s)

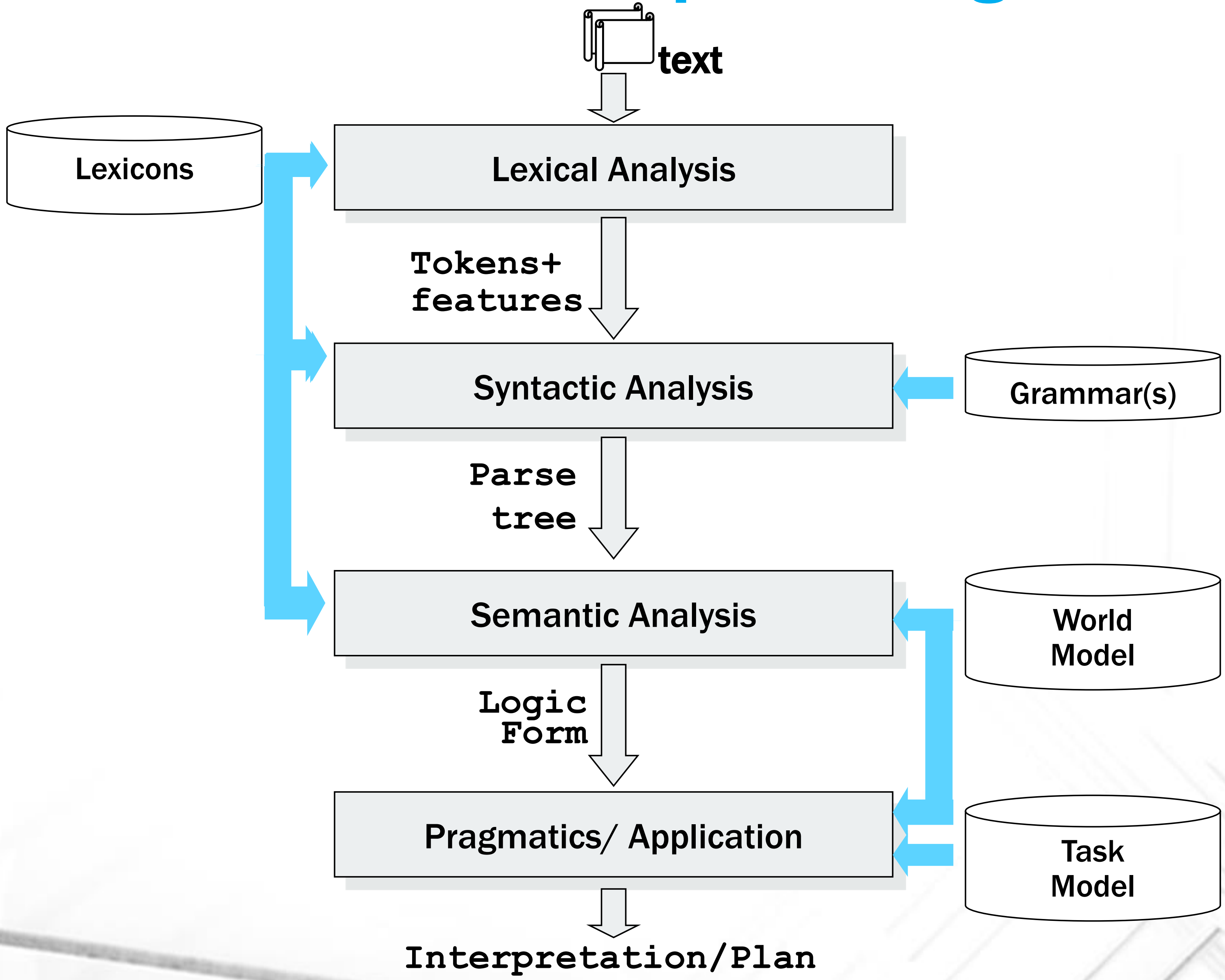
Semantics, Natural Language & Learning: tasks

- From Learning to Read to Knowledge Distillation as a (integrated pool of)

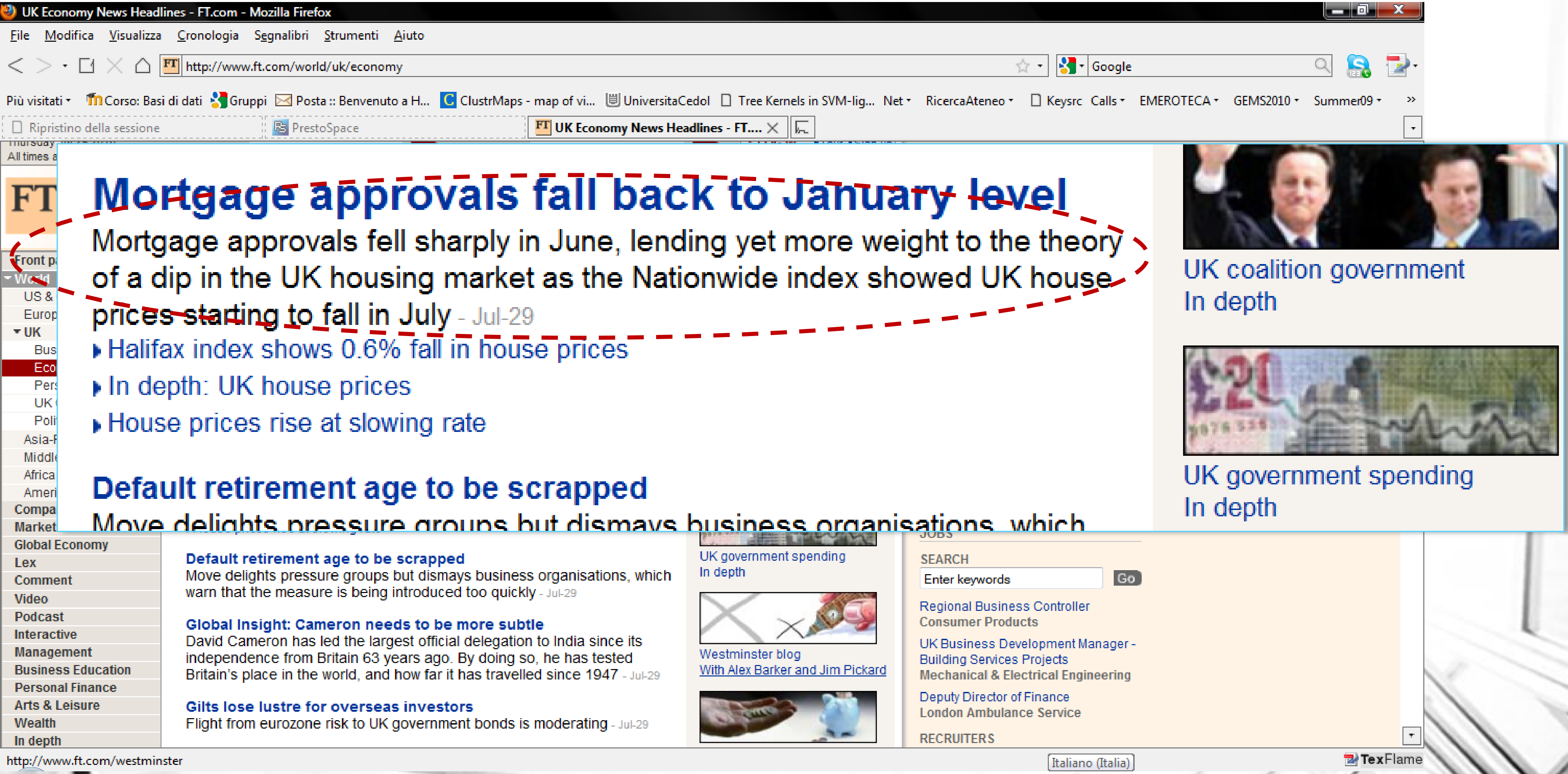
  Semantic interpretation Task(s)

  - **Information Extraction**
    - Entity Recognition and Classification
    - Relation Extraction
    - Semantic Role Labeling (Shallow Semantic Parsing)
  - **Estimation of Text Similarity**
    - Structured Text Similarity/Textual Entailment Recognition
    - Sense disambiguation
  - **Semantic Search, Question Classification and Answer Ranking**
  - **Knowledge Acquisition**, e.g. ontology learning
  - **Social Network Analysis, Opinion Mining**

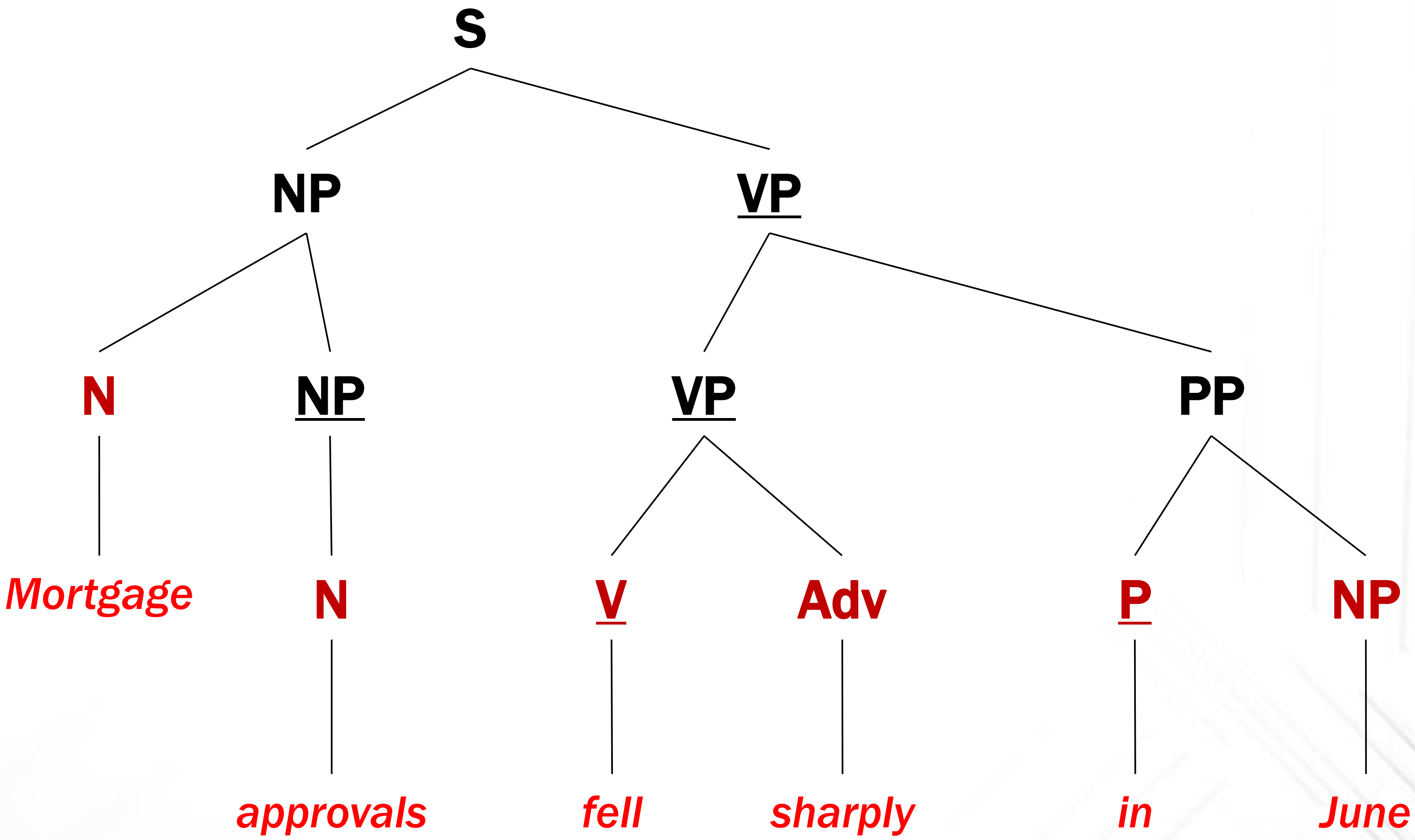
NLP: the standard processing chain

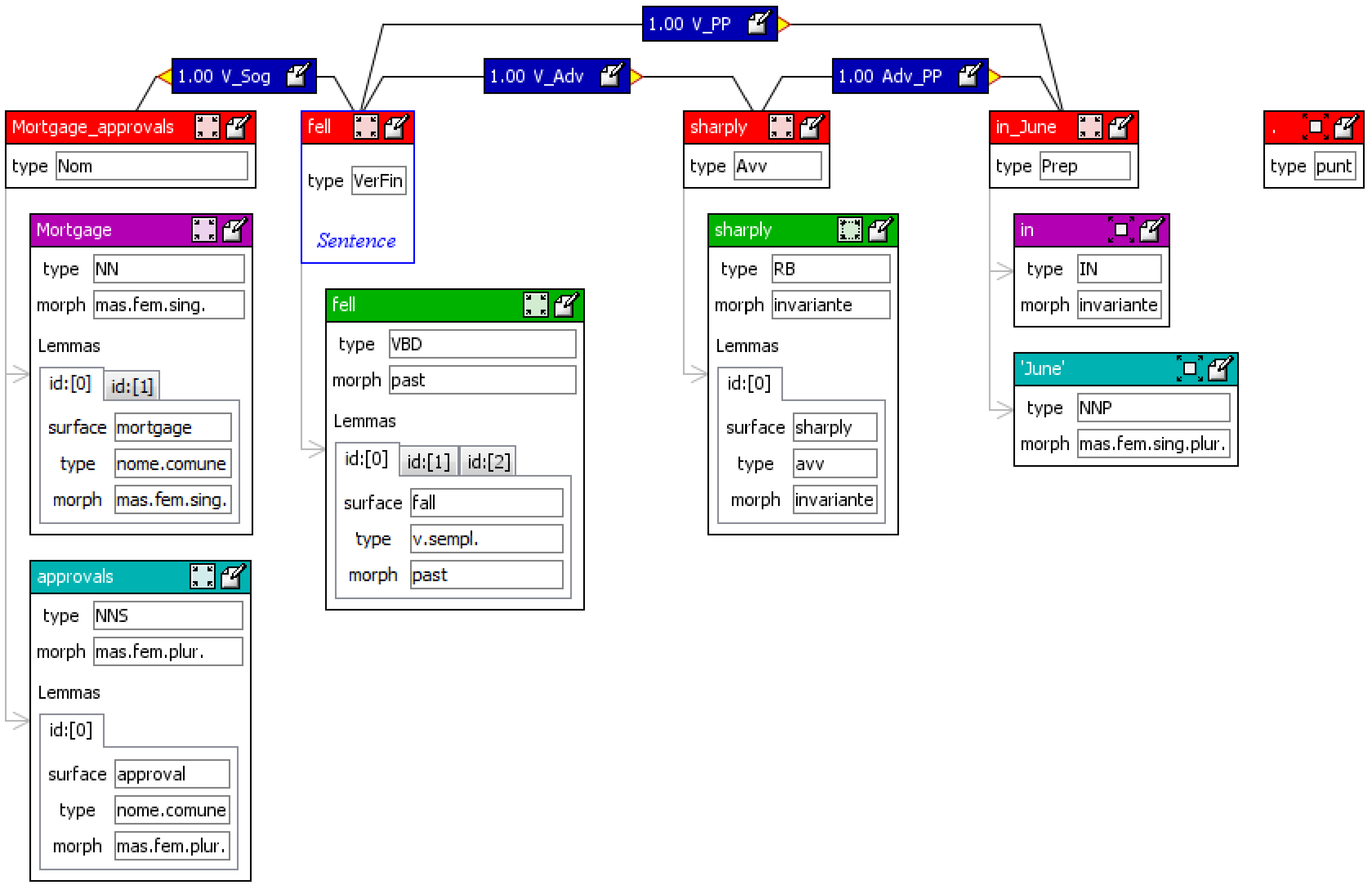


Grammatical Analysis

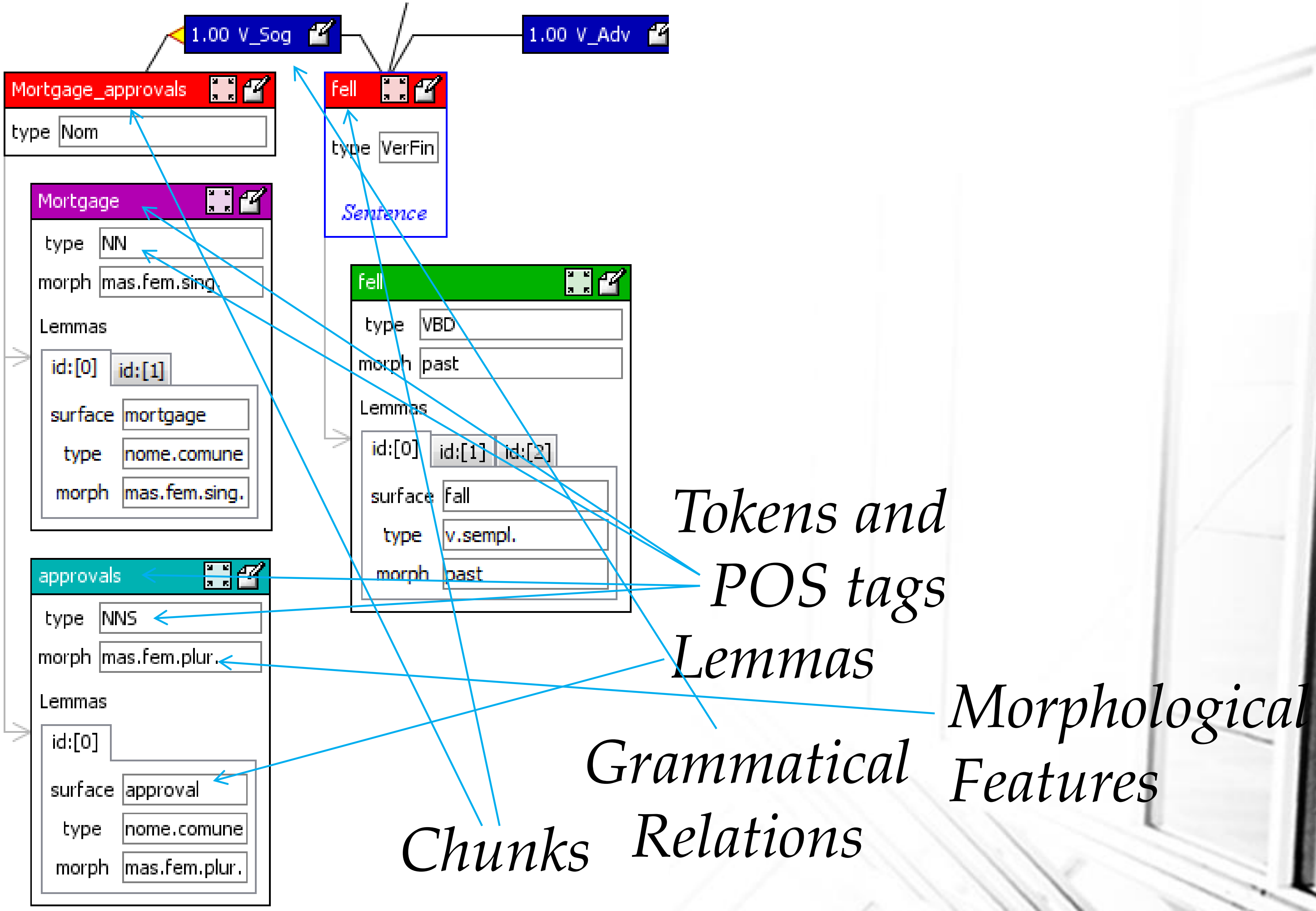


Constituency-based Parsing





FT (July, 29): ***Mortgage approvals fell sharply in June.***



FT (July, 29): *Mortgage approvals fell sharply in June.*

Challenges for Parsing

- Huge complexity as for the ambiguity in the morphosyntactic descriptions of words
  - E.g. La vecchia porta la sbarra
- Interdependency with semantic information
  - Most ambiguity cannot be solved only at the grammatical level
  - Lexical Semantic information is crucial as grammatical structures are constrained by word senses
    - Operating in a market vs. Operating a patient

Semantics

- What is the meaning of the sentence

  John saw Kim?

- Desirable Properties:
  - **It should be derivable as a function of the indivdual constituents,** i.e. the meanings of costituents such as *Kim*, *John* and *see*
  - **Independent from syntactic phenomena**, e.g. *Kim was seen by John* is a paraphrasis
  - It must be directy used to trigger some **inferences:**
    - *Who was seen by John? Kim!*
    - *John saw Kim. He started running to her.*

A Truth conditional semantics

S
saw(s,k)

NP
s

Sam
s

VP
{ x : saw(x,k)}

V
{ <x,y> : saw(x,y)}

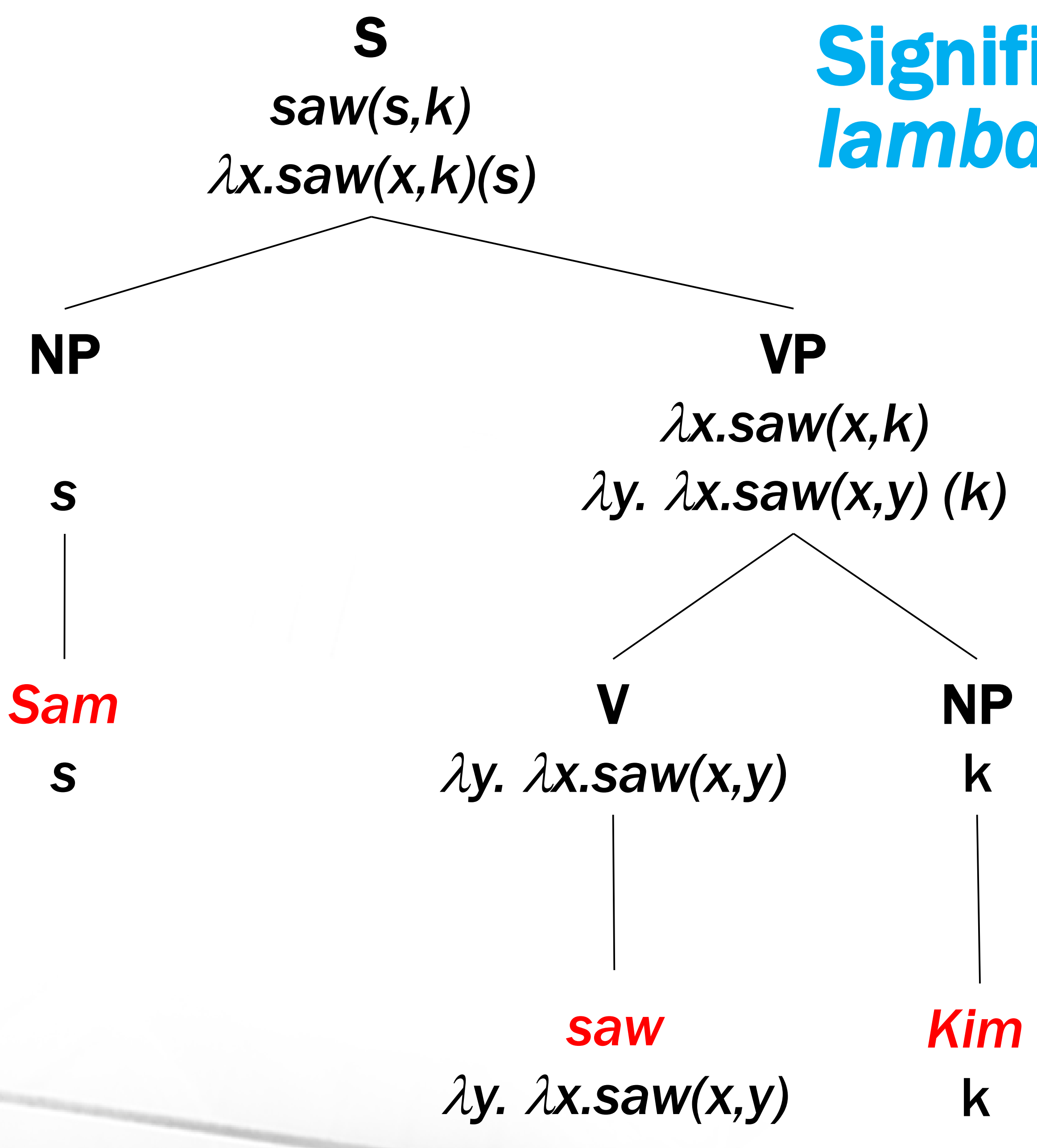
saw
{ <x,y> : saw(x,y)}

NP
k

Kim
k

John saw Kim

Significato come calcolo di *lambda espressioni*



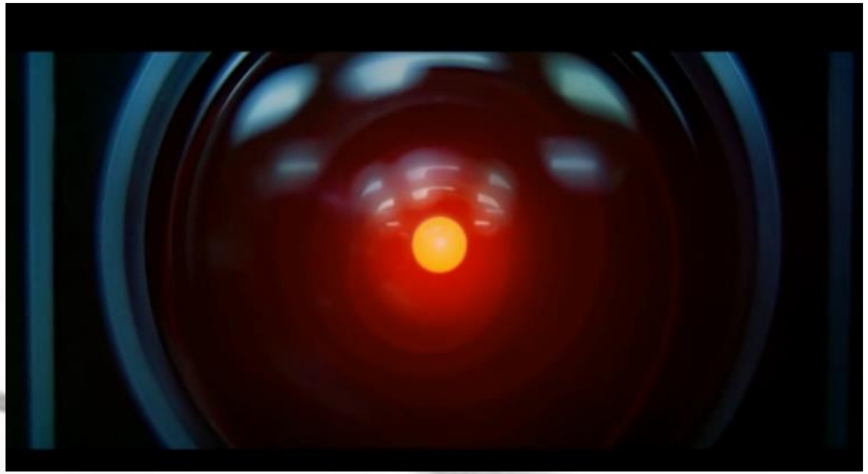
Overview

- Intelligenza Artificiale e Lingue parlate e scritte
  - Informazioni e Rappresentazioni coinvolte
  - Sfide (ri)correnti, battaglie (già) vinte e rischi inerenti …
- Elaborazione Automatica delle Lingue: Modelli, Metodi e *Risultati*
- *break*
- Ruolo delle Tecnologie dell'Apprendimento ed Applicazioni:
  - Sviluppo Automatico di Dizionari, Lessici Semantici ed Ontologie
  - Riconoscimento di fenomeni semantici
  - Trattamento Semantico della Documentazione Investigativa
  - Sistemi Web-based di Opinion Mining, Market Watch & Brand Reputation Management

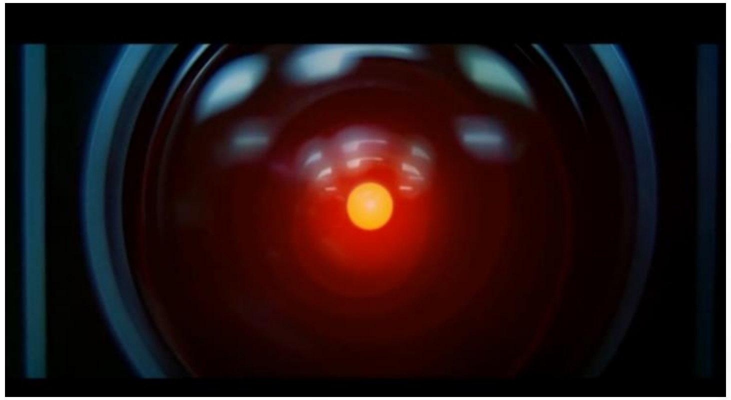
Which Knowledge?

- HAL 9000, da "*2001: A Space Odyssey*"

- Dave: *Open the pod bay doors, Hal*.
- HAL: *I'm sorry Dave, I'm afraid I can't do that*.

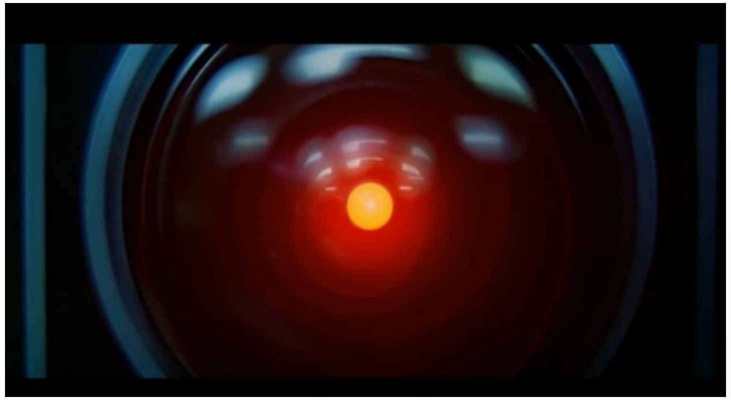


What’s HAL knowledge?



- **Recognition & Synthesis of spoken language**
  - Dictionaries (spelling)
  - Phonetics (how to produce/recognize sound)
- **Understanding**
  - **Lexical Knowledge**
    - What do the words mean?
    - How they combine (`*pod bay door*’)
  - **Knowledge about the syntagmatic structure of sentences**
    - *I’m I do, Sorry that afraid Dave I’m can’t*

What’s HAL knowledge?



- **Dialogue & pragmatics**
  - “*open the door*” is a request (and not a declaration or a search query)
  - Replying is a type of action that imply kindness (even if a planning to kill is in progress …)
  - It is useful to behave cooperatively (*I’m afraid, I can’t…*)
  - What about `*that*’ in `*I can’t do that*’?

Language Processing as a (*semantic*) *interpretation process*

- Processing a text corresponds to understand a number of aspects related to its *meaning*
  - Thematic Domain (e.g. science/housekeeping/economics)
  - Operational Objectives (e.g. e-mail *spam*)
  - Involved Entitites, such as people or locations
  - Potential events described (e.g. facts told by news)
  - Obbiettivi comunicativi (e.g. dialogue, orders/declarations/planning)
- Outcome: an explicit *representation of the text meaning* ...
- able to trigger different inferences
  (e.g. IR *relevance, planning*, knowledge updates, ...)

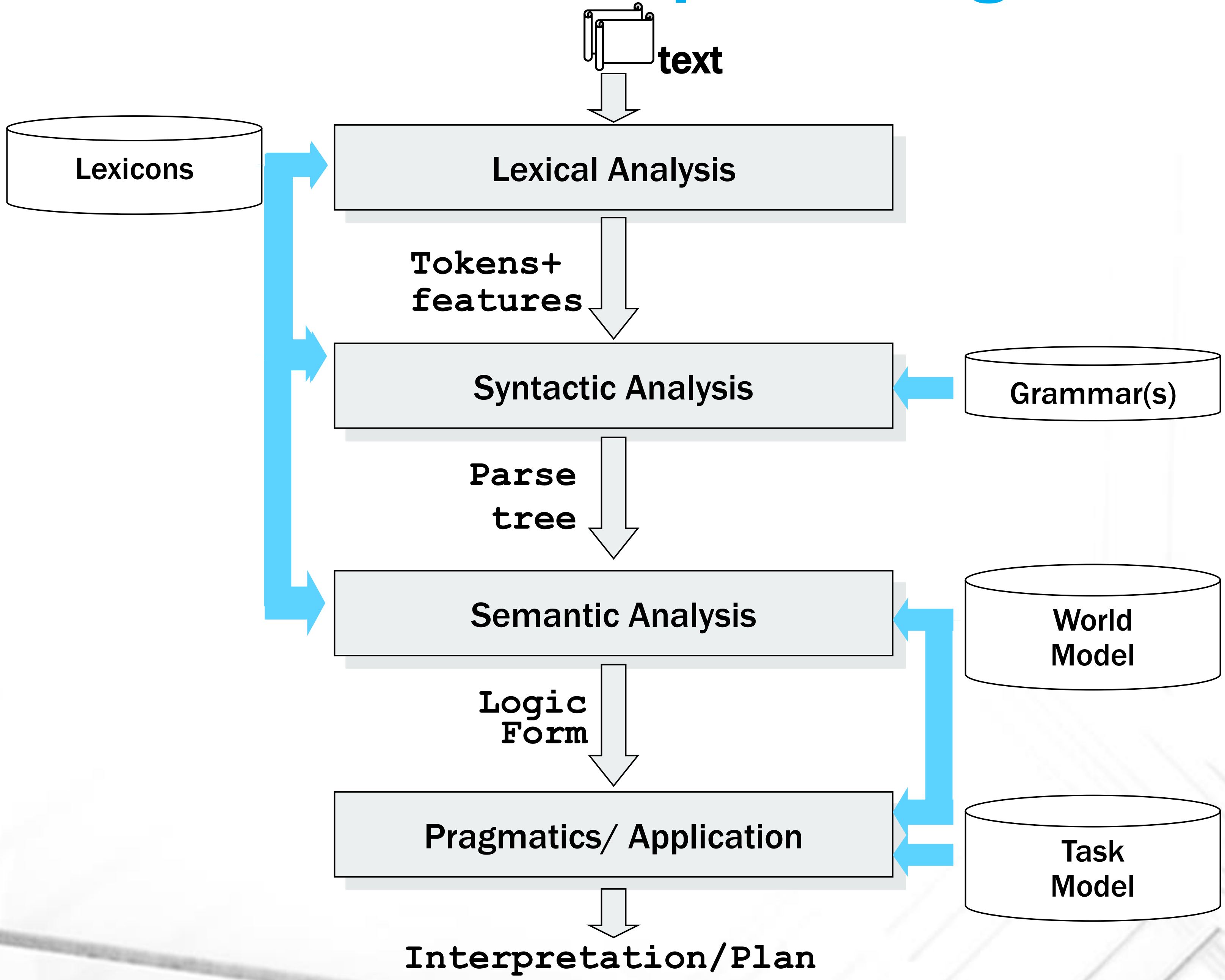
Some Reflections

- Understading *linguistic information* requires specific knowledge about:
  - The natural language itself (e.g. *grammar*)
  - The world (e.g. *bay door*, *Dave* or *opening*)
  - How language make **reference** to the world
- NLP applications deals with texts by exploiting the specific context:
  - Application purposes, e.g. document search
  - The domain and the operational context of an application
  - The distinction between language producer (speaker/writer) and consumer (hearer/reader)

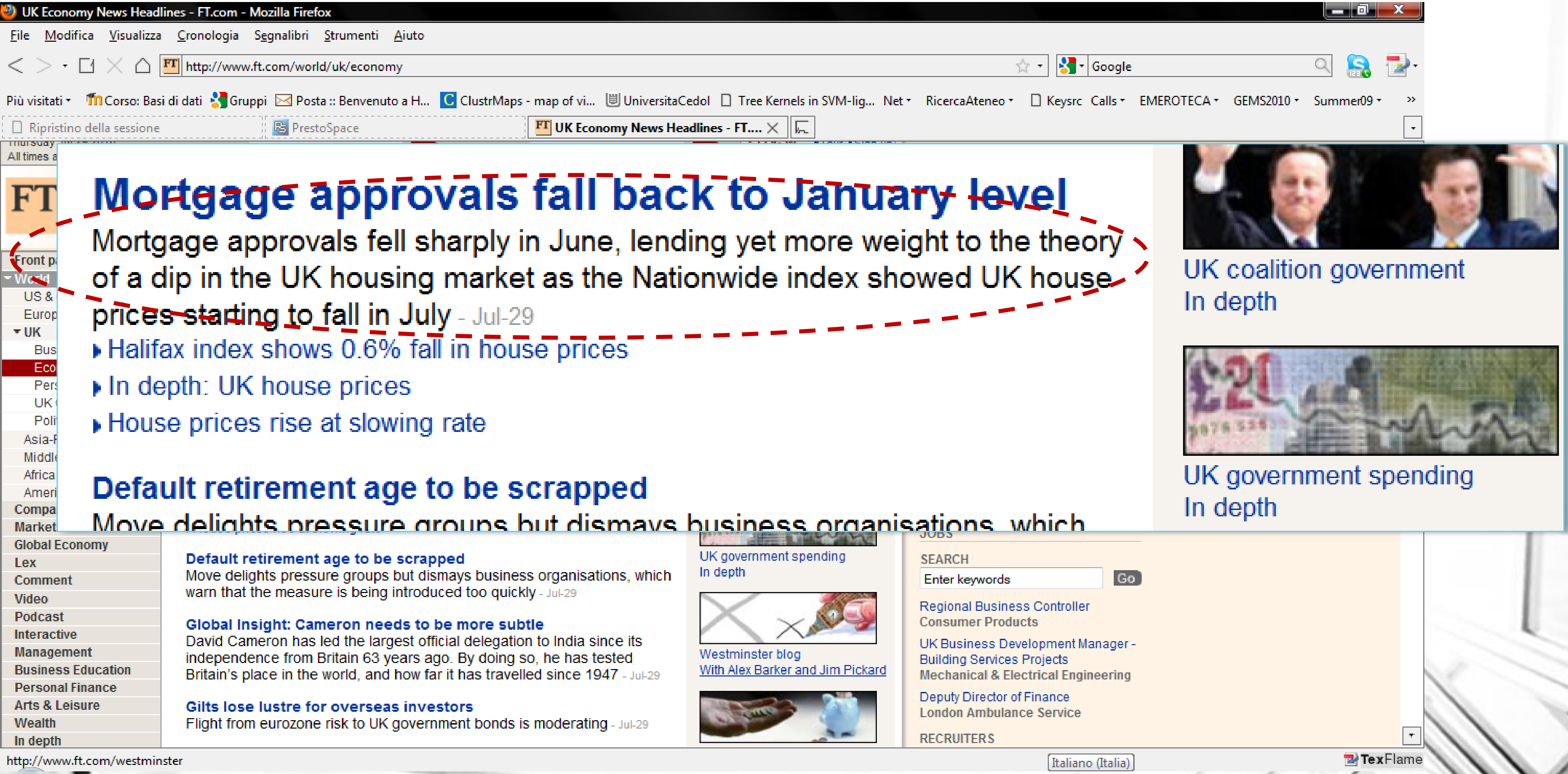
Major Challenges

- *Linguistic Accuracy* in approximating the human-level of performance
- *Robustness* (errors/noise/incompleteness)
- *Scale*
  - Coverage of the phenomena (Lexicons/Grammars)
- *Expressivity*
  - Dictionaries, Lexicons and *Thesauri*
  - World Models and types of inference
- *Flexibility*
  - Adequate performance across linguistic variability (e.g. producer vs. consumer)
- *Naturalness*

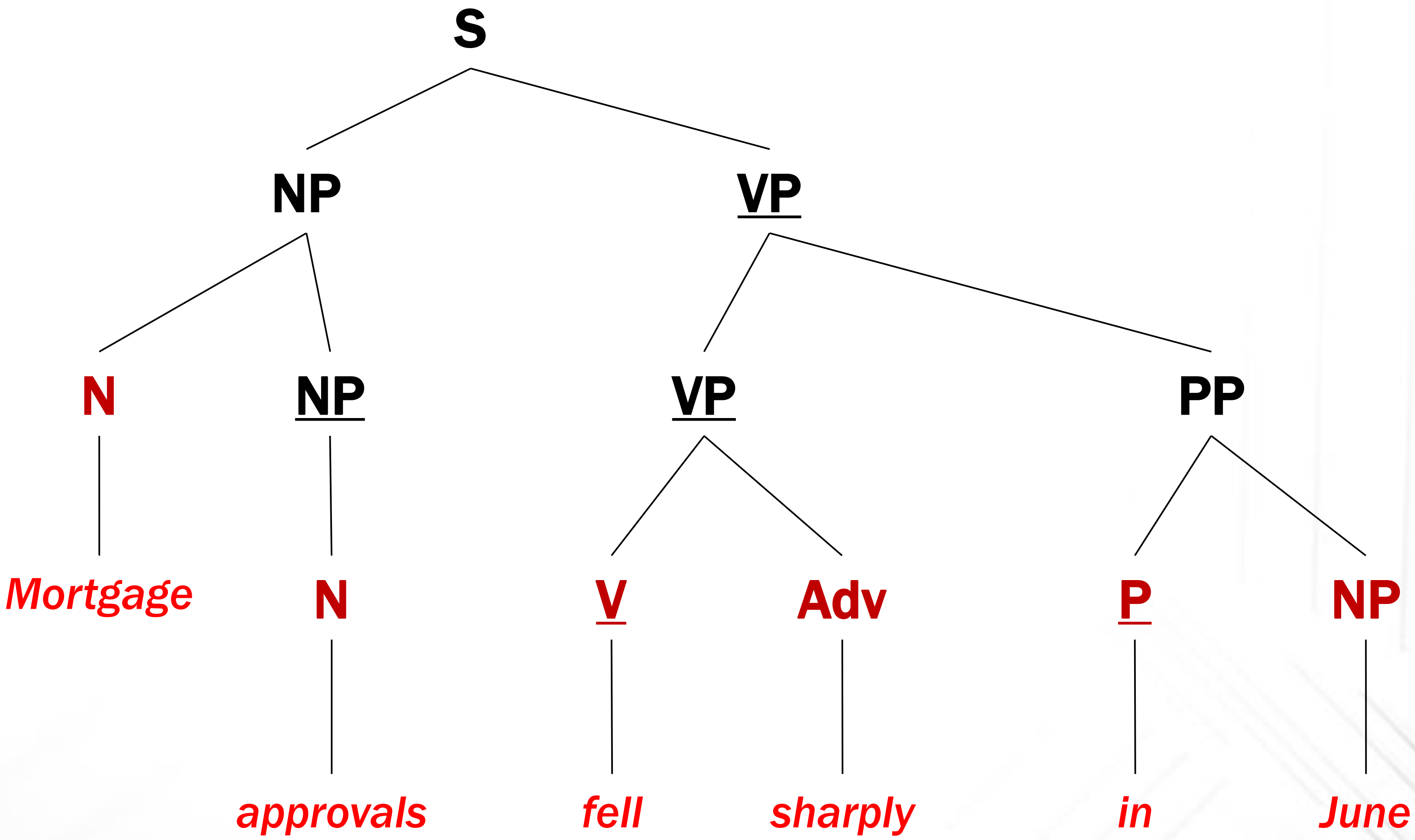
NLP: the standard processing chain



Grammatical Analysis



Constituency-based Parsing



Challenges for Parsing

- Huge complexity as for the ambiguity in the morphosyntactic descriptions of words
  - E.g. La vecchia porta la sbarra
- Interdependency with semantic information
  - Most ambiguity cannot be solved only at the grammatical level
  - Lexical Semantic information is crucial as grammatical structures are constrained by word senses
    - Operating in a market vs. Operating a patient

Semantics

- What is the meaning of the sentence

  John saw Kim?

- Desirable Properties:
  - **It should be derivable as a function of the indivdual constituents,** i.e. the meanings of costituents such as *Kim*, *John* and *see*
  - **Independent from syntactic phenomena**, e.g. *Kim was seen by John* is a paraphrasis
  - It must be directy used to trigger some **inferences:**
    - *Who was seen by John? Kim!*
    - *John saw Kim. He started running to her.*

A Truth conditional semantics

S
saw(s,k)

NP
s

Sam
s

VP
$\{ x : saw(x,k) \}$

V
$\{ \langle x,y \rangle : saw(x,y) \}$

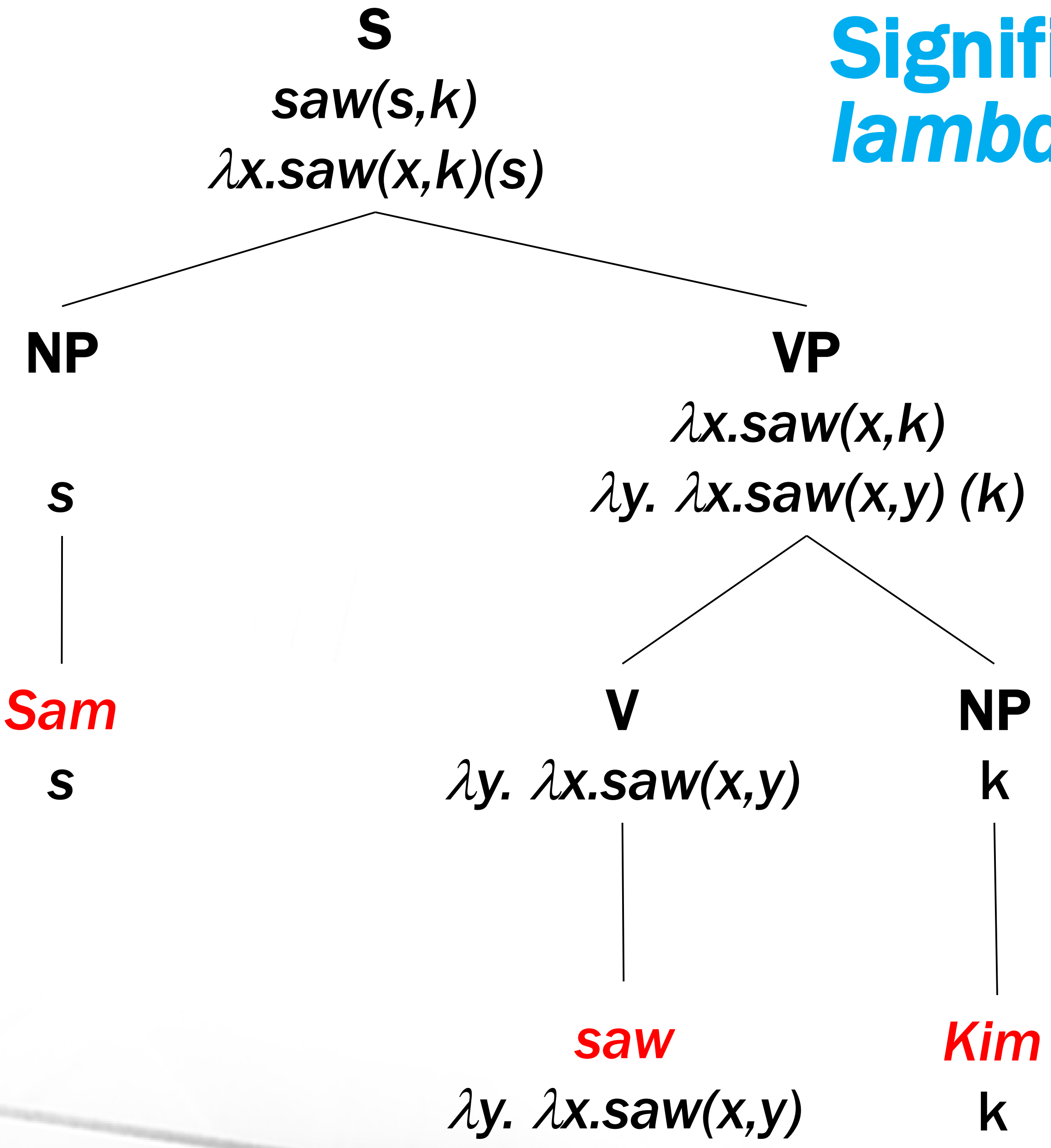
saw
$\{ \langle x,y \rangle : saw(x,y) \}$

NP
k

Kim
k

John saw Kim

Significato come calcolo di *lambda espressioni*

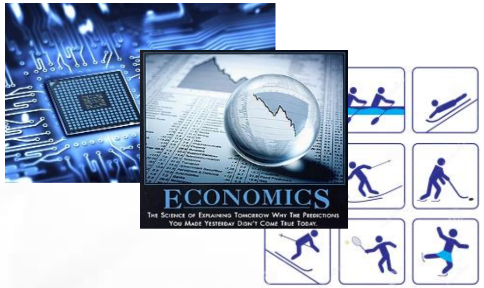


Fenomeni Semantici di interesse

- Entità descritte nei testi (persone, luoghi, organizzazioni, date, espressioni numeriche o monetarie)

- Relazioni / Associazioni tra entità

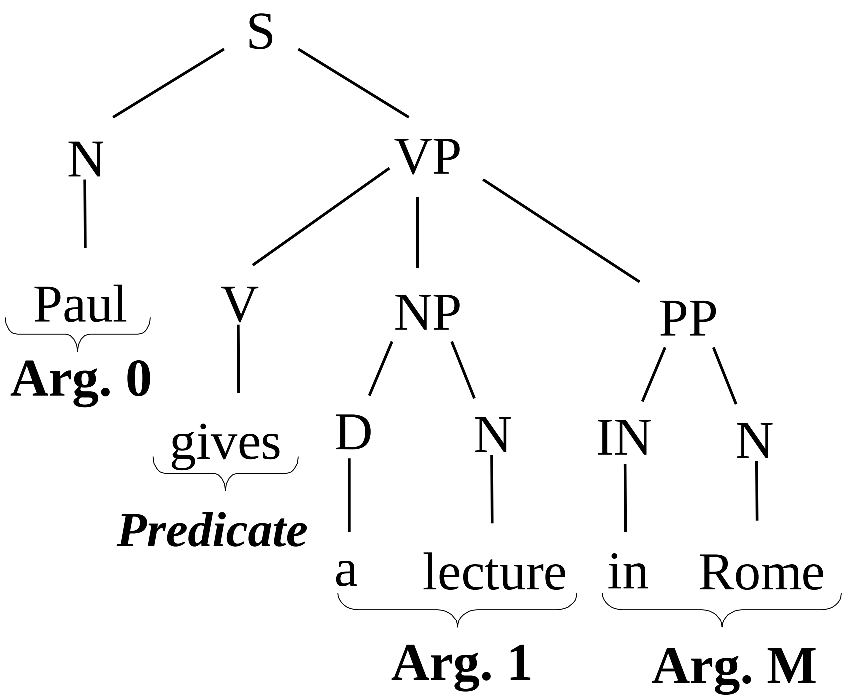
- Fatti ed Eventi



- Temi / Topic / Contesto / Dominio

Predicazione ed Argomenti

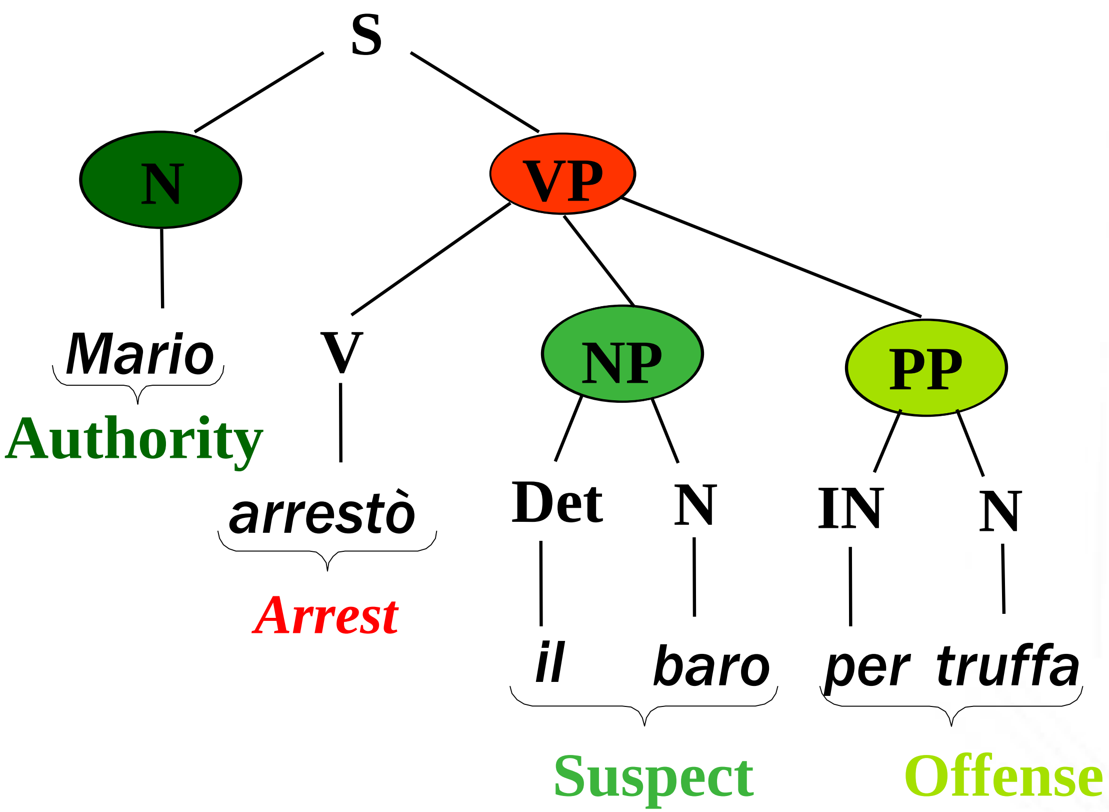
- Il *mapping* sintassi-semantica



Annotazioni Semantiche diverse: PropBank vs. FrameNet

Linking syntax to semantics: Semantic Role Labeling

Mario arrestò il baro per truffa



[Il baro]$_{\text{Suspect}}$ *[fu arrestato]*$_{\text{Arrest}}$ *[da Mario]*$_{\text{Authority}}$ *[per truffa]*$_{\text{Offense}}$

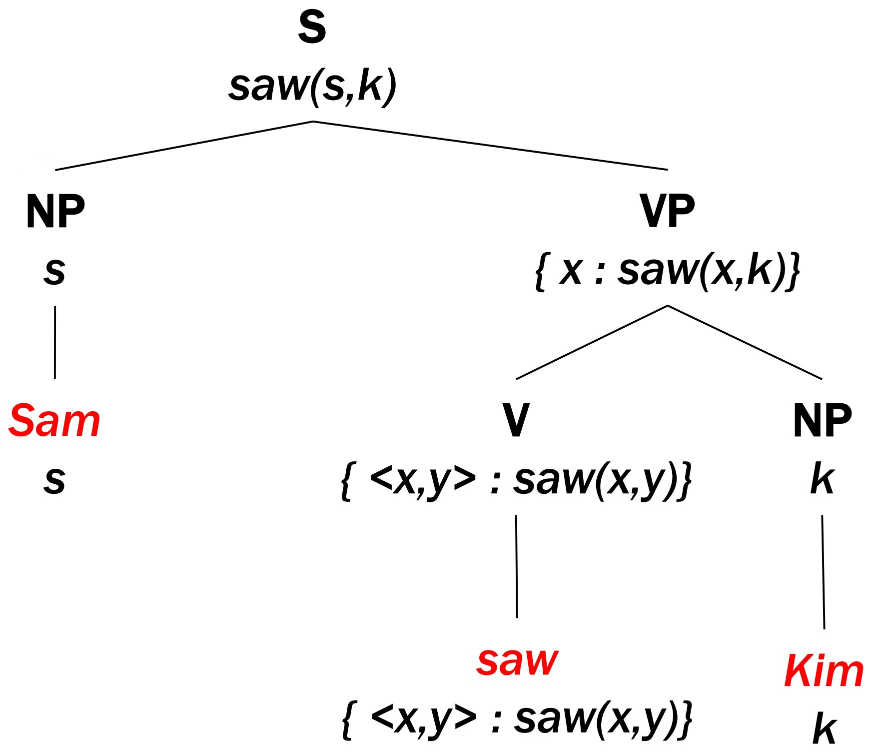
Semantics

- For the sentence:

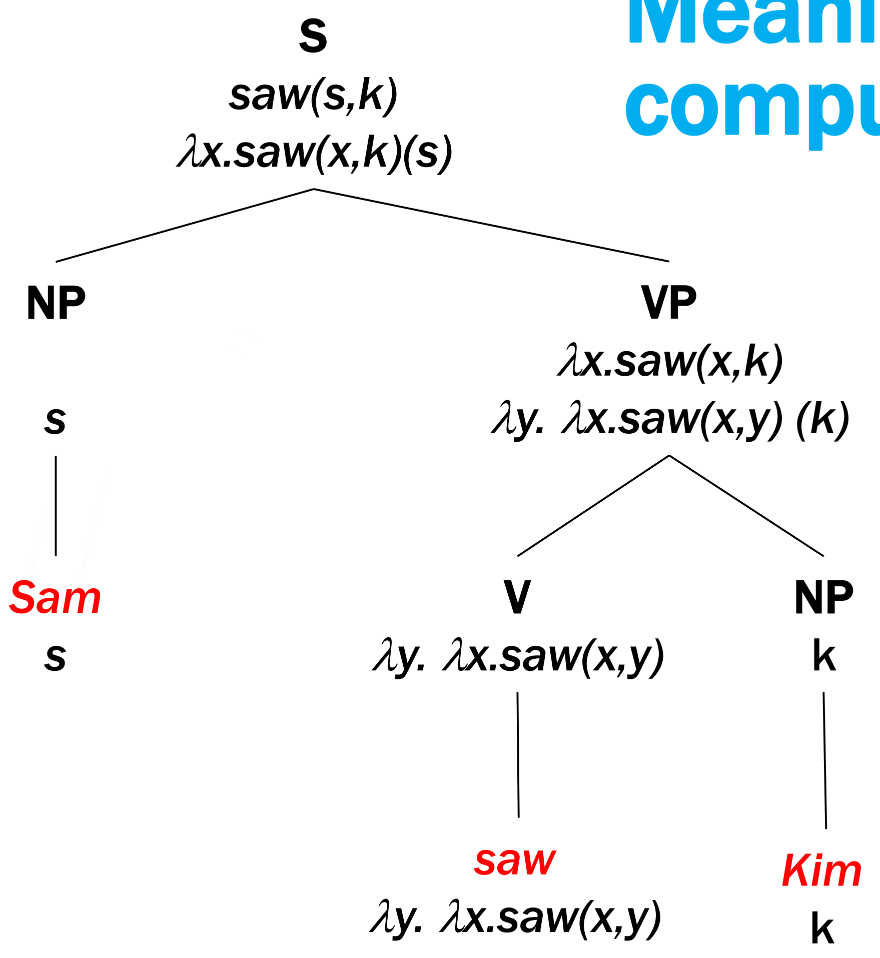
  John saw Kim

- What abut its **meaning?**
- Properties:
  - **It must be derivable compositionally**, i.e. from the meanings of the individual constituents, i.e. ***Kim***, ***John*** and **see**
  - **Independence on syntactic phenomenon**, e.g. *Kim was seen by John*
  - It must **support inferences**
    - *Who was seen by John?*
    - *John saw Kim. He started running to her.*

Truth conditional view on meaning



Meaning as a computation



Syntax and Semantics in textual data

- Compositionality
- The meaning of a complex expression is solely determined by the meanings of its constituent expressions and the rules used to combine them.
- "*I will consider a language to be a set (finite or infinite) of sentences, each finite in length and constructed out of a finite set of elements. All natural languages are languages in this sense. Similarly, the set of "sentences" of some formalized system of mathematics can be considered a language*" Chomsky 1957

Syntax

- In linguistics, **syntax** is the study of the rules that govern the structure of sentences, and which determine their relative grammaticality.
- Such rules govern a number of language phenomena as systems for phonology, morphology, syntax as well as discourse

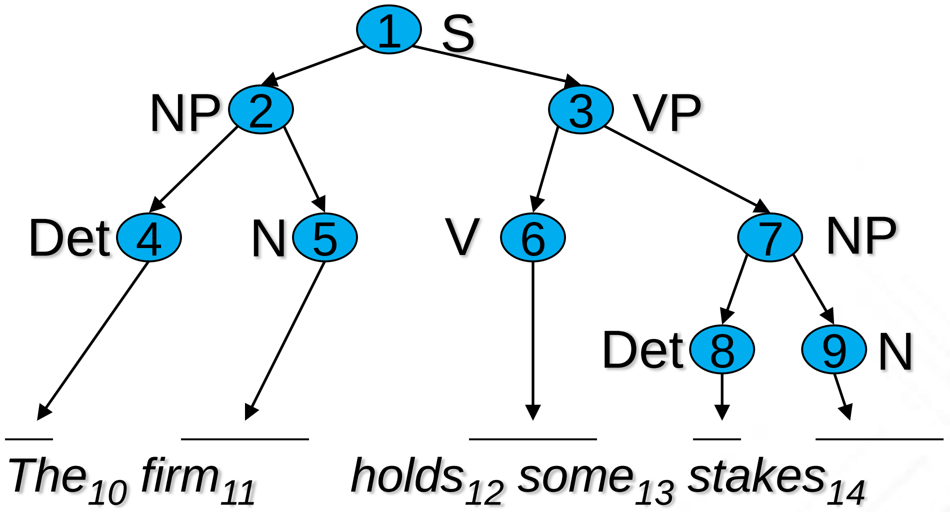
Parse Trees

- The representation of the parsing result is a structure that expresses:
  - The order of constituent elements in the sentence
  - The grammatical type of constituents
  - The hierarchical organization of constitunts
- The structures able to express these properties are the derivation trees also called parse trees

Grammars and Trees

"The firm holds some stakes"

- Vn={S,NP,VP,Det,N},   Axiom: S
- Productions: {S⟶NP VP, VP⟶V NP, NP⟶Det N}
- Derivation:
  - S > NP VP > Det N VP > *The* N VP > *The firm* VP > *The firm* V NP > *The firm holds* NP > *The firm holds* Det N > *The firm holds some* N > *The firm holds some stakes*



Semantics: sense, predicates and arguments

- In the traditional view grammatical categories give rise at the semantic level to 0-arity, unary (e.g. nouns) or n-ary (e.g. verbs) predicates
- Sentence semantics is expressed via quantified logical formulas
- E.g. *John gives mary the book*
- `Give( John, Mary, book)`
- `∃e1,e2,e3:`

```
give(e1, e2, e3) ∧ book(e3) ∧
                        name(e1, John) ∧
                              name(e2, Mary)
```

Semantics

- Words *senses* activates predicates
  - Bank/money vs. bank/river
  - bank_1(X) vs. bank_2(X)
- Verbal predicates express
  - Events/states
  - Relation among participants

- See unit "Ambiguity and Variability in Natural Languages " on the Course Web page
- For a discussion about a Prolog-based approach see "Semantic Analysis in Prolog "

Three Perspectives on Meaning

- Lexical Semantics
  - The meanings of individual words
- Formal Semantics (or Compositional Semantics or Sentential Semantics)
  - How those meanings combine to make meanings for individual sentences or utterances
- Discourse or Pragmatics
  - How those meanings combine with each other and with other facts about various kinds of context to make meanings for a text or discourse
  - Dialog or Conversation is often lumped together with Discourse

Lexical Semantic:
Relationships between word meanings

- Homonymy
- Polysemy
- Synonymy
- Antonymy
- Hypernomy
- Hyponomy
- Meronomy

Homonymy

- **Homonymy:**
  - Lexemes that share a form
    - Phonological, orthographic or both
  - But have unrelated, distinct meanings
  - Clear example:
    - Bat (wooden stick-like thing) vs
    - Bat (flying scary mammal thing)
    - Or bank (financial institution) versus bank (riverside)
  - Can be also homophones, homographs, or both:
    - Homophones:
      - *Write* and *right*
      - *Piece* and *peace*

Polysemy

- The bank is constructed from red brick
  I withdrew the money from the bank
- Are those the same sense?
- Or consider the following WSJ example
  - While some banks furnish sperm only to married women, others are less restrictive

  - Which sense of bank is this?
    - Is it distinct from (homonymous with) the river bank sense?
    - How about the savings bank sense?

Metaphor and Metonymy

- Specific types of polysemy
- Metaphor:
  - Germany will pull Slovenia out of its economic slump.
  - *I spent 2 hours on that homework*.
- Metonymy
  - The White House announced yesterday.
  - This chapter talks about part-of-speech tagging
  - Bank (building) and bank (financial institution)

Synonyms

- Word that have the same meaning in some or all contexts.
  - *filbert / hazelnut*
  - *couch / sofa*
  - *big / large*
  - *automobile / car*
  - *vomit / throw up*
  - *Water / H_2O*
- Two lexemes are synonyms if they can be successfully substituted for each other in all situations
  - If so they have the same **propositional meaning**

Synonyms

- But there are few (or no) examples of perfect synonymy.
  - Why should that be?
  - Even if many aspects of meaning are identical still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.
- Example:
  - Water and H_2O
  - I would not say

    I like fresh H_2O after the tennis

Some terminology

- Lemmas and wordforms
  - A **lexeme** is an abstract pairing of meaning and form
  - A **lemma** or **citation form** is the grammatical form that is used to represent a **lexeme**.
    - ***Carpet*** is the lemma for ***carpets, Dormir*** is the lemma for ***duermes*.**
  - Specific surface forms *carpets, sung, duermes* are called **wordforms**

- The lemma *bank* has two **senses:**
  - Instead, a **bank** can hold the investments in a custodial account in the client’s name
  - But as agriculture burgeons on the east **bank**, the river will shrink even more.

- A **sense** is a discrete representation of one aspect of the meaning of a word

Synonymy is a relation between senses rather than words

- Consider the words *big* and *large*
- Are they synonyms?
  - How **big** is that plane?
  - Would I be flying on a **large** or small plane?
- How about here:
  - Miss Nelson, for instance, became a kind of **big** sister to Benjamin.
  - ?Miss Nelson, for instance, became a kind of **large** sister to Benjamin.
- Why?
  - *big* has a sense that means being older, or grown up
  - *large* lacks this sense

Antonyms

- Senses that are opposites with respect to one feature of their meaning
- Otherwise, they are very similar!
  - dark / light
  - short / long
  - hot / cold
  - up / down
  - in / out
- More formally: antonyms can
  - define a binary opposition or opposite ends of a scale (*long/short, fast/slow*)
  - Be **reversives**: *rise/fall, up/down*

Hyponymy

- One sense is a **hyponym** of another if the first sense is more specific, denoting a subclass of the other
  - *car* is a hyponym of *vehicle*
  - *dog* is a hyponym of *animal*
  - *mango* is a hyponym of *fruit*
- Conversely
  - *vehicle* is a hypernym/superordinate of *car*
  - *animal* is a hypernym of *dog*
  - *fruit* is a hypernym of *mango*

superordinate	vehicle	fruit	furniture	mammal
hyponym	car	mango	chair	dog

Hypernymy more formally

- Extensional:
  - The class denoted by the superordinate extensionally includes the class denoted by the hyponym
- Entailment:
  - A sense A is a hyponym of sense B if being an A entails being a B
- Hyponymy is usually transitive
  - (A hypo B and B hypo C entails A hypo C)

II. WordNet

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary
  - Versions for other languages are under development

Category	Unique Forms
Noun	117,097
Verb	11,488
Adjective	22,141
Adverb	4,601

WordNet

- Home page:
  - http://www.cogsci.princeton.edu/cgi-bin/webwn

Format of Wordnet Entries

The noun "bass" has 8 senses in WordNet.

1. bass^1 - (the lowest part of the musical range)
2. bass^2, bass part^1 - (the lowest part in polyphonic music)
3. bass^3, basso^1 - (an adult male singer with the lowest voice)
4. sea bass^1, bass^4 - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass^1, bass^5 - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass^6, bass voice^1, basso^2 - (the lowest adult male singing voice)
7. bass^7 - (the member with the lowest range of a family of musical instruments)
8. bass^8 - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

The adjective "bass" has 1 sense in WordNet.

1. bass^1, deep^6 - (having or denoting a low vocal or instrumental range) *"a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"*

WordNet Noun Relations

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	$breakfast^1 \rightarrow meal^1$
Hyponym	Subordinate	From concepts to subtypes	$meal^1 \rightarrow lunch^1$
Member Meronym	Has-Member	From groups to their members	$faculty^2 \rightarrow professor^1$
Has-Instance		From concepts to instances of the concept	$composer^1 \rightarrow Bach^1$
Instance		From instances to their concepts	$Austen^1 \rightarrow author^1$
Member Holonym	Member-Of	From members to their groups	$copilot^1 \rightarrow crew^1$
Part Meronym	Has-Part	From wholes to parts	$table^2 \rightarrow leg^3$
Part Holonym	Part-Of	From parts to wholes	$course^7 \rightarrow meal^1$
Antonym		Opposites	$leader^1 \rightarrow follower^1$

WordNet Verb Relations

Relation	Definition	Example
Hypernym	From events to superordinate events	$fly^9 \rightarrow travel^5$
Troponym	From a verb (event) to a specific manner elaboration of that verb	$walk^1 \rightarrow stroll^1$
Entails	From verbs (events) to the verbs (events) they entail	$snore^1 \rightarrow sleep^1$
Antonym	Opposites	$increase^1 \Longleftrightarrow decrease^1$

WordNet Hierarchies

```
Sense 3
bass, basso --
(an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, someone...
               => organism, being
                  => living thing, animate thing,
                     => whole, unit
                        => object, physical object
                           => physical entity
                              => entity
               => causal agent, cause, causal agency
                  => physical entity
                     => entity

Sense 7
bass --
(the member with the lowest range of a family of
musical instruments)
=> musical instrument, instrument
   => device
      => instrumentality, instrumentation
         => artifact, artefact
            => whole, unit
               => object, physical object
                  => physical entity
                     => entity
```

How is “*sense*” defined in WordNet?

- The set of near-synonyms for a WordNet sense is called a **synset** (**synonym set**); it’s their version of a sense or a concept
- Example: **chump** as a noun to mean
  - ‘a person who is gullible and easy to take advantage of’

{chump^1, fool^2, gull^1, mark^9, patsy^1, fall guy^1, sucker^1, soft touch^1, mug^2}

- Each of these senses share this same gloss
- Thus for WordNet, the meaning of this sense of **chump** *is* this list.

Word Similarity

- Synonymy is a binary relation
  - Two words are either synonymous or not
- We want a looser metric
  - Word similarity or
  - Word distance
- Two words are more similar
  - If they share more features of meaning

Word Similarity

- Actually these are really relations between **senses**:
  - Instead of saying "*bank is like fund*"
  - We say
    - Bank1 *is similar to* fund3
    - Bank2 *is similar to* slope5
- Similarity are computed over both words and senses

Why word similarity

- Spell Checking
- Information retrieval
- Question answering
- Machine translation
- Natural language generation
- Language modeling
- Automatic essay grading

Syntactic Argument Structures

- (Verbal) Relations require a fixed number of participants, called arguments
- The syntactic structure predicts the number and type of arguments through subcategorization frames
  - (Bob (gave (Mary) (the book) (on Monday)))
  - (Bob (gave (the book) (to Mary) (on Monday)))

Thematic roles

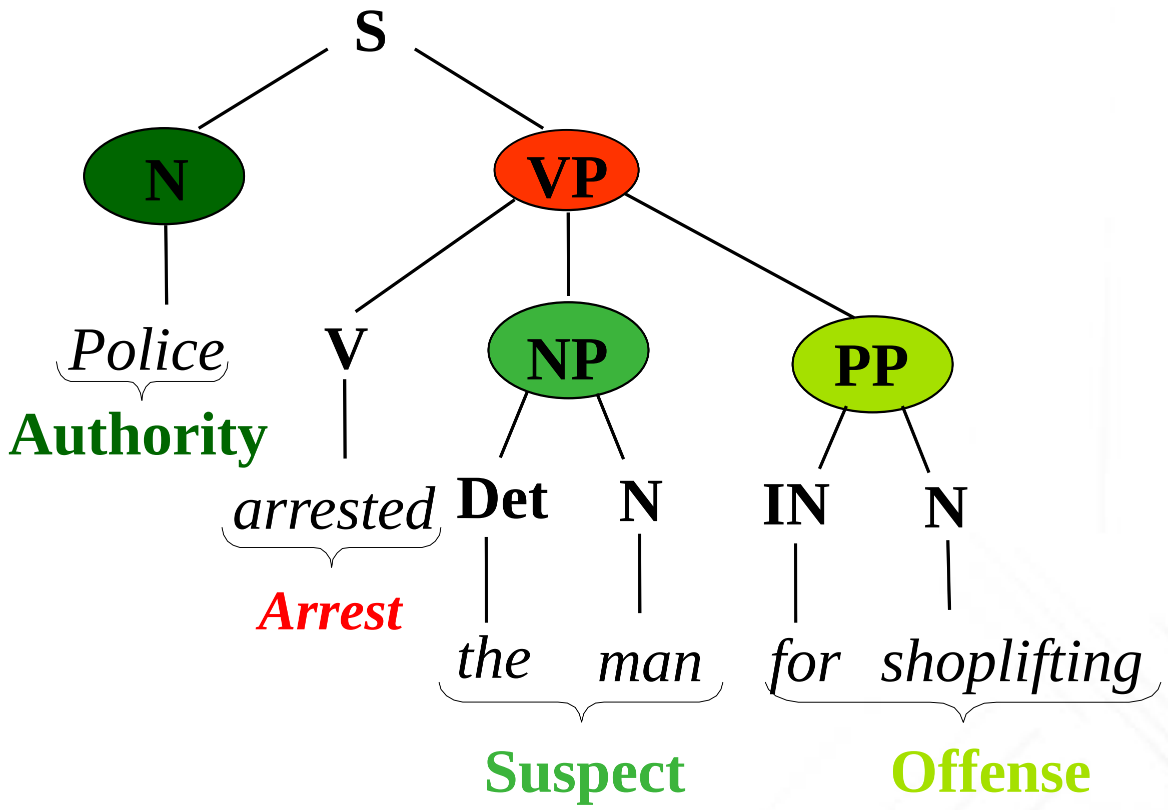
- Arguments play specific roles, called thematic roles, depending on the predicate but invariant across different syntactic structures giving rise to predicate argument structures
  - `give(Agent:` *Bob*, `Theme:` *the_book*, `Recipient:` *Mary*`)`

- Thematic roles of individual arguments <u>are indexed</u> by their predicates

- *General* and *lexicalized* roles have been introduced

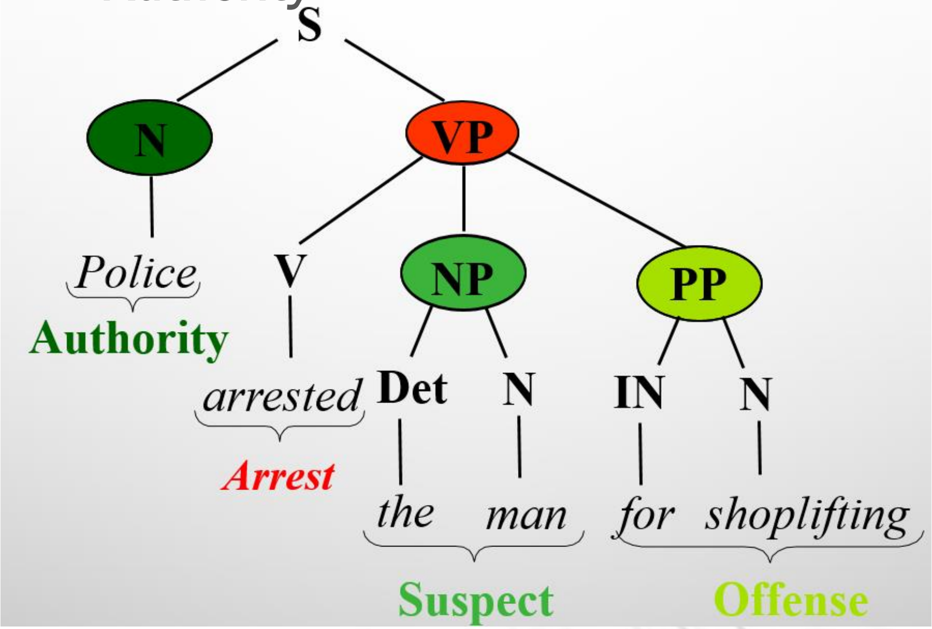
Linking syntax to semantics

- *Police arrested the man for shoplifting*



A tabular vision

Word	Predicate	Semantic Role
Police	-	Authority
arrested	Target	Arrest
the	-	SUSPECT
man	-	SUSPECT
for	-	OFFENSE
Shoplifting	-	OFFENSE



Semantics in NLP: Resources

- Lexicalized Models
  - Propbank
  - NomBank
- Framenet
  - Inspired by frame semantics
  - Frames are lexicalized prototoypes for real -world situations
  - Participants are called frame elements (roles)

Frame Semantics

- Research in Empirical Semantics suggests that **words represents categories of experience** (*situations*)
- A ***frame*** is a cognitive structuring device (i.e. a kind of prototype) indexed by *words* and used to support understanding (Fillmore, 1975)
  - *Lexical Units* **evoke** a Frame in a sentence
- Frames are made of *elements* that express participants to the situation (*Frame Elements*)
- During communication LUs evoke the frames

Fram

Frame: KILLING		
A KILLER or CAUSE causes the death of the VICTIM.		
Frame Elements	KILLER	**John** drowned Martha.
	VICTIM	John drowned **Martha**.
	MEANS	The flood exterminated the rats **by cutting off access to food**.
	CAUSE	**The rockslide** killed nearly half of the climbers.
	INSTRUMENT	It's difficult to suicide **with only a pocketknife**.
Predicates	annihilate.v, annihilation.n, asphyxiate.v,assassin.n, assassinate.v, assassination.n, behead.v, beheading.n, blood-bath.n, butcher.v, butchery.n, carnage.n, crucifixion.n, crucify.v, deadly.a, decapitate.v, decapitation.n, destroy.v, dispatch.v, drown.v, eliminate.v, euthanasia.n, euthanize.v, …	

Frame Semantics

- Lexical descriptions are expected to define the indexed frame and the frame elements with their realization at the syntactic level:
  - *John bought a computer from Janice for 1000 $*
- Mapping into syntactic arguments
  - the buyer is (usually) in the subject position
- Obligatory vs. optional arguments
- Selectional preferences
  - The seller and the buyer are usually “humans” or “social groups”

The FrameNet project

- The aims
  - Create a lexical resource by describing a significant portion of English in terms of precise and rich frame semantics
- The output
  - Frame Database: a structured system of Frames and Fes
  - Lexical database: syntactic and semantic descriptions of frame-evoking words (N,V,A)
  - Annotated Corpus: wide coverage examples

FrameReport - Mozilla Firefox

Frame Report (recent data)

| Top of Frame Index | Top of Lexical Unit Index |

Committing_crime

Definition:

A Perpetrator (generally intentionally) commits a Crime, i.e. does something not permitted by the laws of society.

They PERPETRATED a felony by substituting a lie for negotiations.

The suspect had allegedly COMMITTED the crime to gain the attention of a female celebrity.

FEs:

Core:

Crime [Cr] An act, generally intentional, that has been formally forbidden by law.

How can he COMMIT treason against the King of England in a foreign country , if he is not English?

He PERPETRATED a crime against mother nature.

Perpetrator [Perp] The individual that commits a Crime.

How can he COMMIT treason against the King of England in a foreign country , if he is not English?

He PERPETRATED a crime against mother nature.

Non-Core:

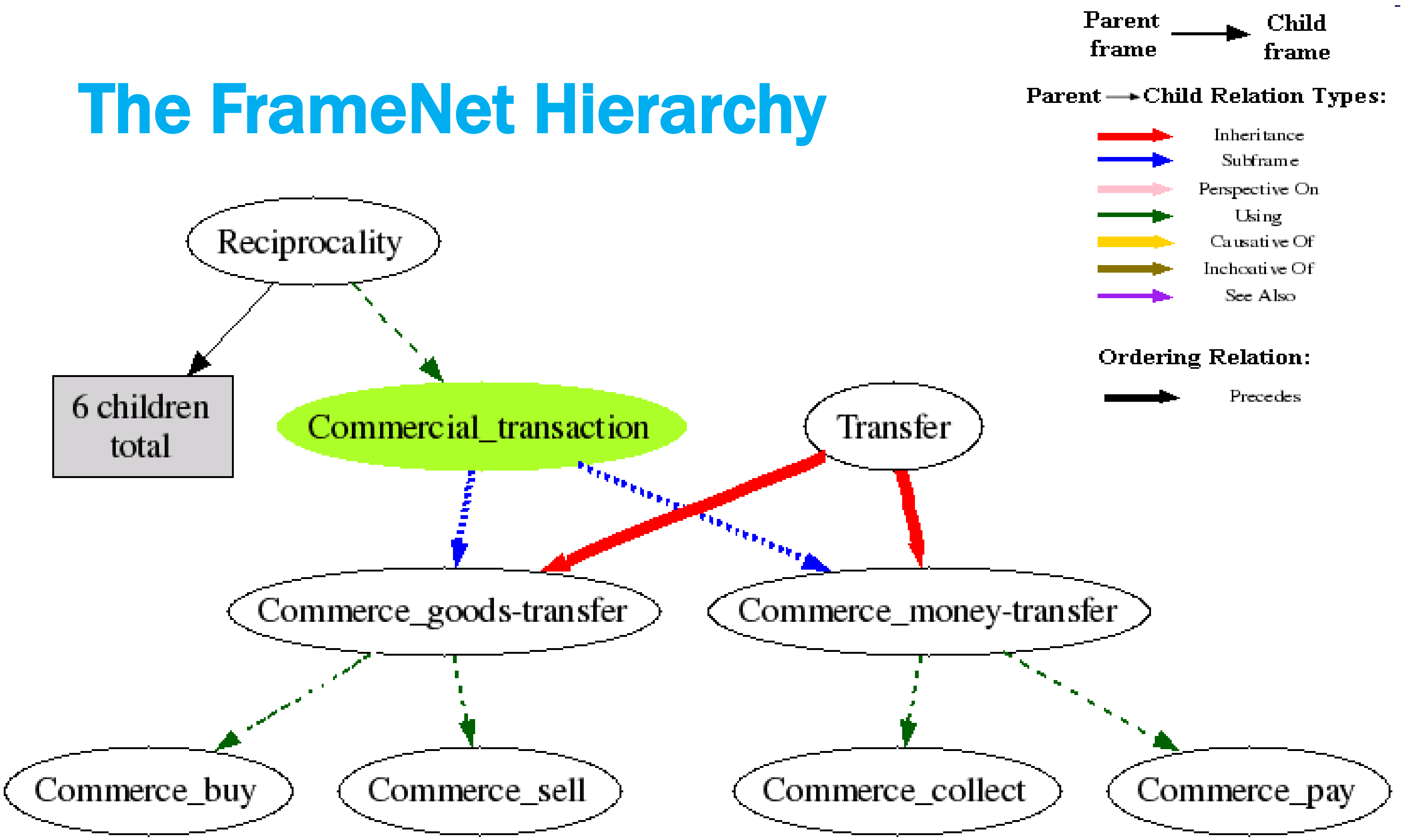
Frequency [Freq] The frequency with which a Crime is committed.

The average serial killer COMMITS a crime every five years.

Instrument [Inst] The Instrument used in committing the crime.

Most crimes are COMMITTED with a firearm.

The FrameNet Hierarchy



Framenet - Data

- Methodology of constructing FrameNet
  - Define/discover/describe frames
  - Decide the participants (frame elements)
  - List lexical units that evoke the frame
  - Find example sentences in the BNC and annotate them
- Corpora
  - FrameNet I -British National Corpus only
  - FrameNet II -LDC North American Newswire corpora
- Size
  - >10,000 lexical units, >825 frames, >135,000 sentences
- http://framenet.icsi.berkeley.edu

Using Framenet

- See later in the slides: Semantic Role Labeling

Overview

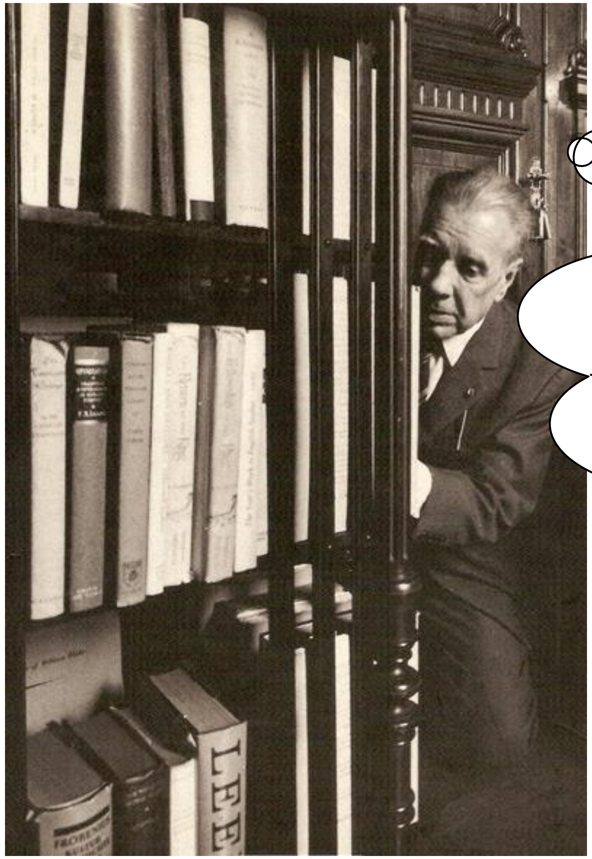
- Intelligenza Artificiale e Lingue parlate e scritte
    - Informazioni e Rappresentazioni coinvolte
    - Sfide (ri)correnti, battaglie (già) vinte e rischi inerenti …
- Elaborazione Automatica delle Lingue: Modelli, Metodi e *Risultati*
- break

- Ruolo delle Tecnologie dell'Apprendimento ed Applicazioni:
    - Sviluppo Automatico di Dizionari, Lessici Semantici ed Ontologie
    - Trattamento Semantico della Documentazione Investigativa
    - Sistemi Web-based di Opinion Mining, Market Watch & Brand Reputation Management

Overview

- Intelligenza Artificiale e Lingue parlate e scritte
  - Informazioni e Rappresentazioni coinvolte
  - Sfide (ri)correnti, battaglie (già) vinte e rischi inerenti …
- Elaborazione Automatica delle Lingue: Modelli, Metodi e *Risultati*
- *break*
- Ruolo delle Tecnologie dell'Apprendimento ed Applicazioni:
  - Sviluppo Automatico di Dizionari, Lessici Semantici ed Ontologie
  - Riconoscimento di fenomeni semantici
  - Trattamento Semantico della Documentazione Investigativa
  - Sistemi Web-based di Opinion Mining, Market Watch & Brand Reputation Management

Il Linguaggio come sistema di regole



*… comincia qui la mia disperazione di scrittore. Ogni **linguaggio** è **un alfabeto di simboli** il cui uso presuppone **un passato che gli interlocutori condividono; come trasmettere agli altri l’infinito Aleph** che la mia timorosa memoria a stento abbraccia?*

(*) J.L.Borges, “*L’aleph*”, 1949.

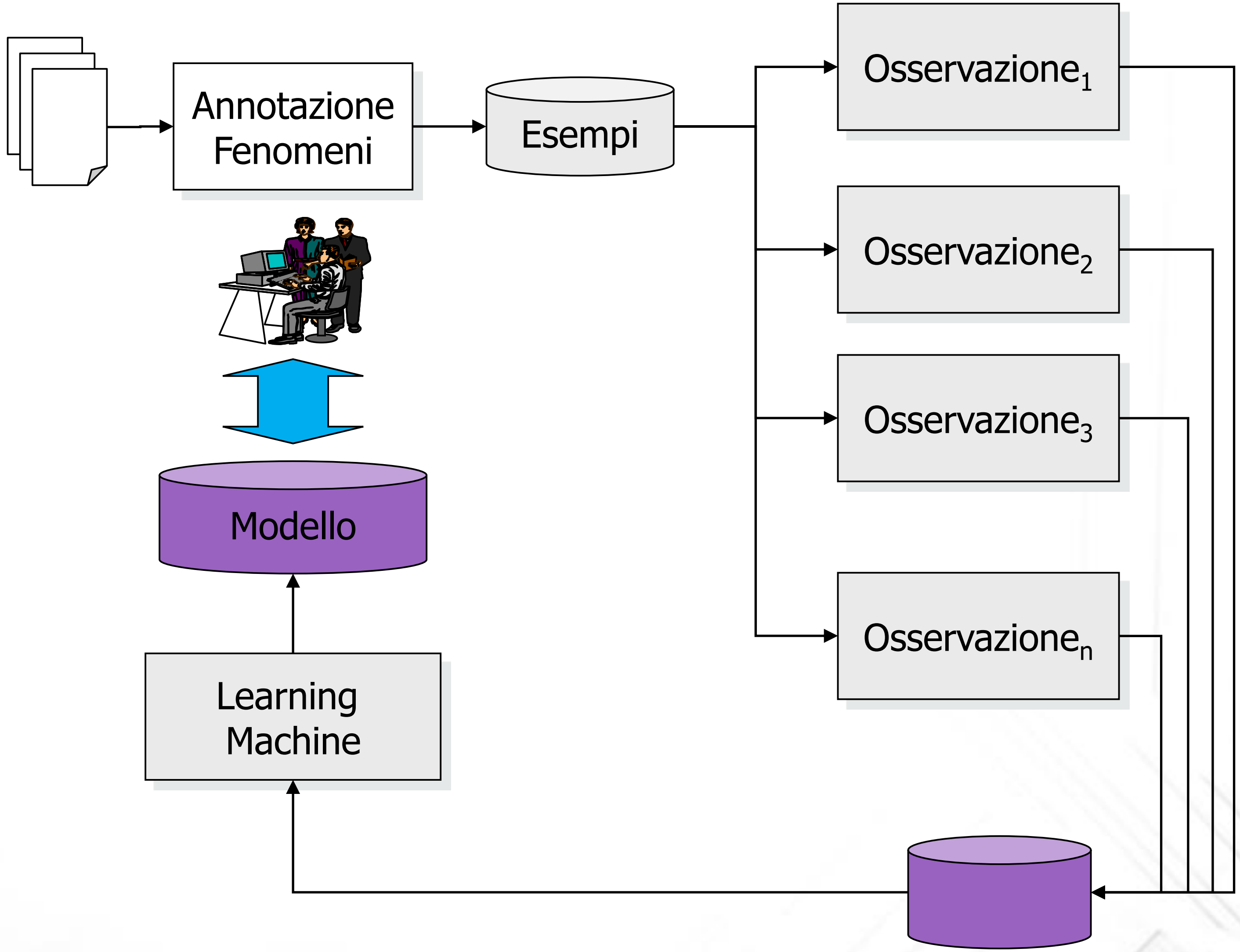
- … Il significato può essere appreso e riconosciuto nelle prassi del suo uso quotidiano
  - *The meaning of a word is to be defined by the rules for its use, not by the feeling that attaches to the words*
    L. Wittgenstein's Lectures, Cambridge 1932-1935.

- Riconoscere un significato consiste nel mappare un testo ad una esperienza (prassi) attraverso meccanismi quali **la analogia** o la approssimazione di **funzioni di equivalenza** o la **minimizzazione del rischio di sbagliare**

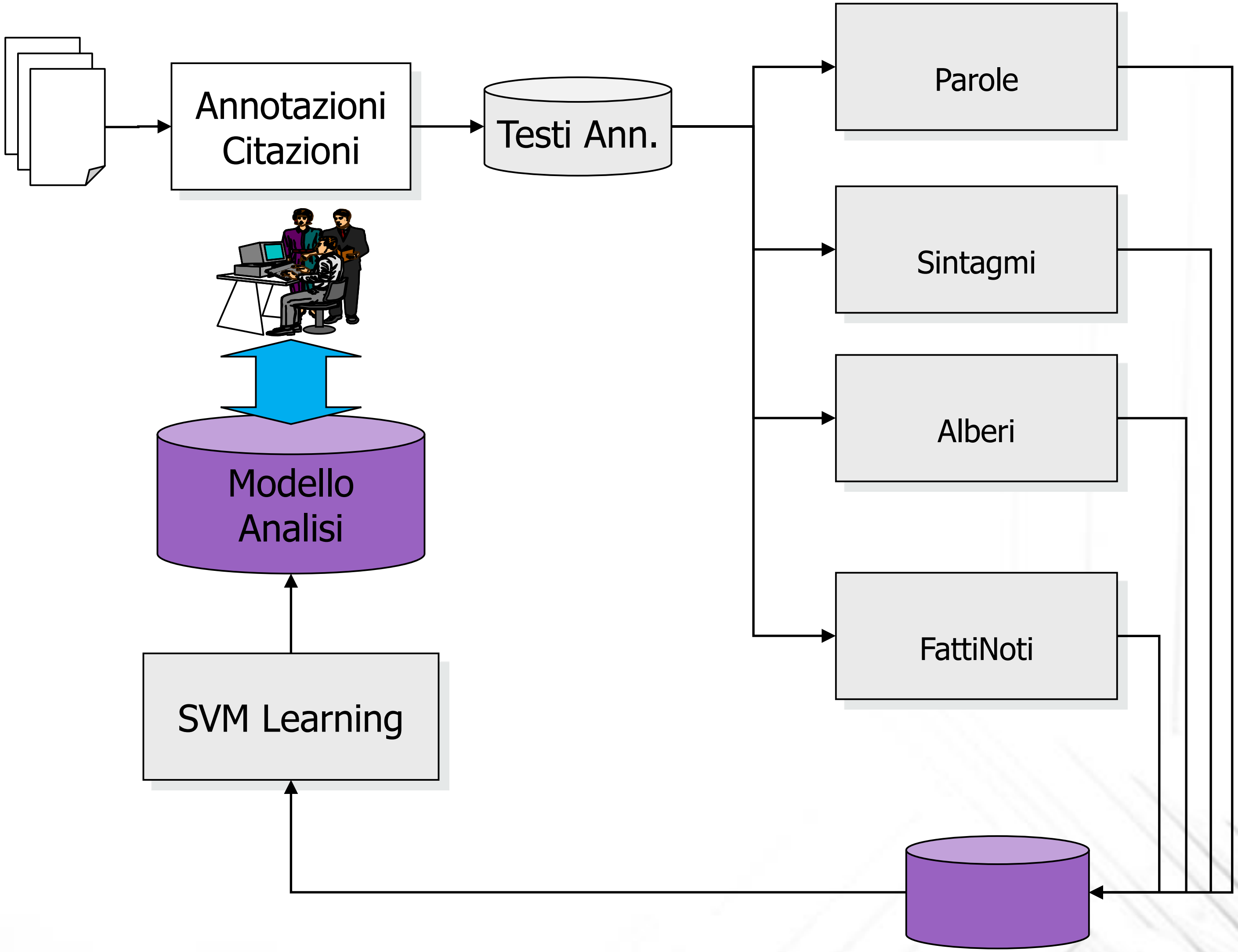
- L'intepretazione si trasforma dunque nella ***induzione di una funzione di decisione a partire dall'esperienza***

Una prospettiva differente

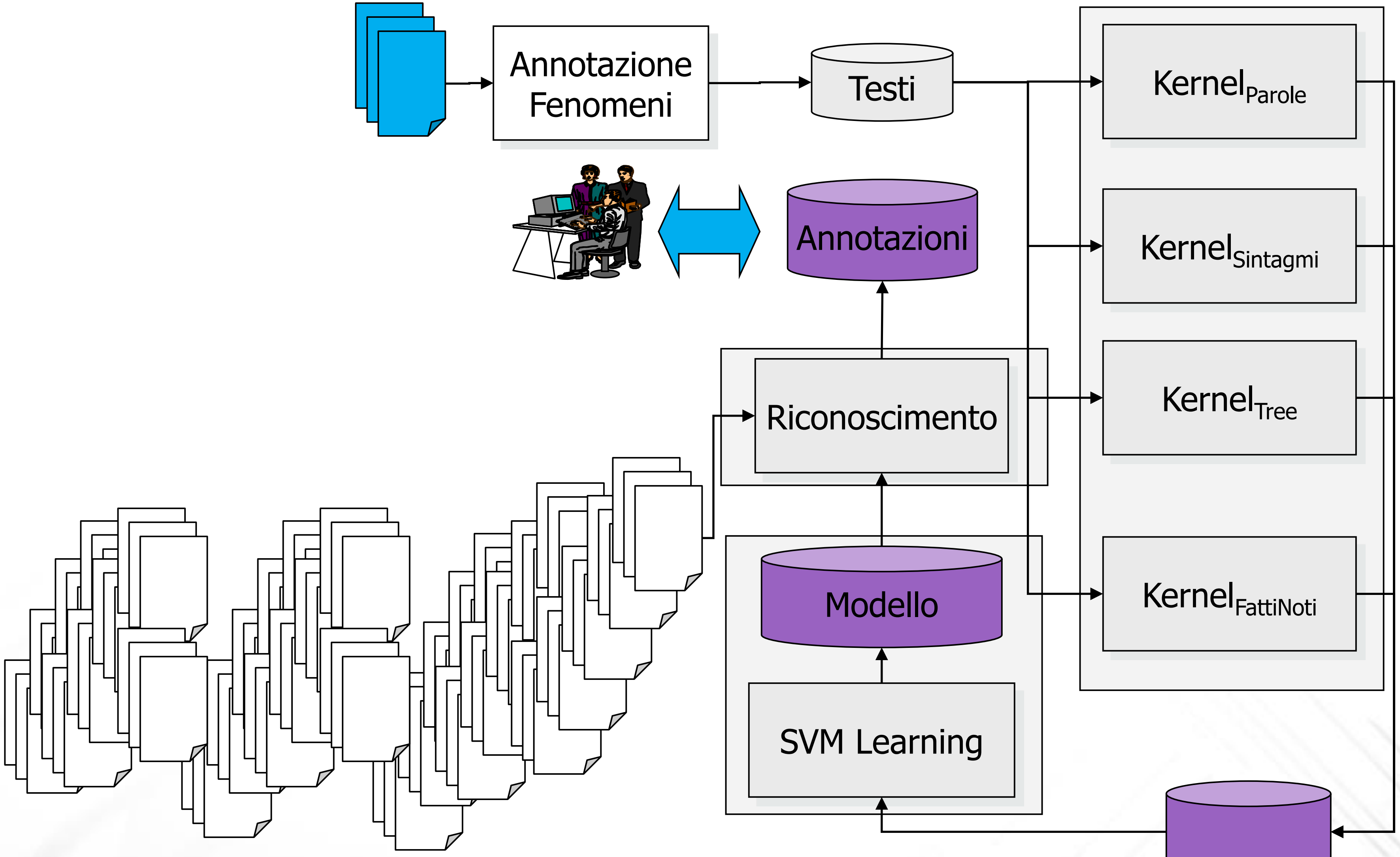
Un Processo Induttivo



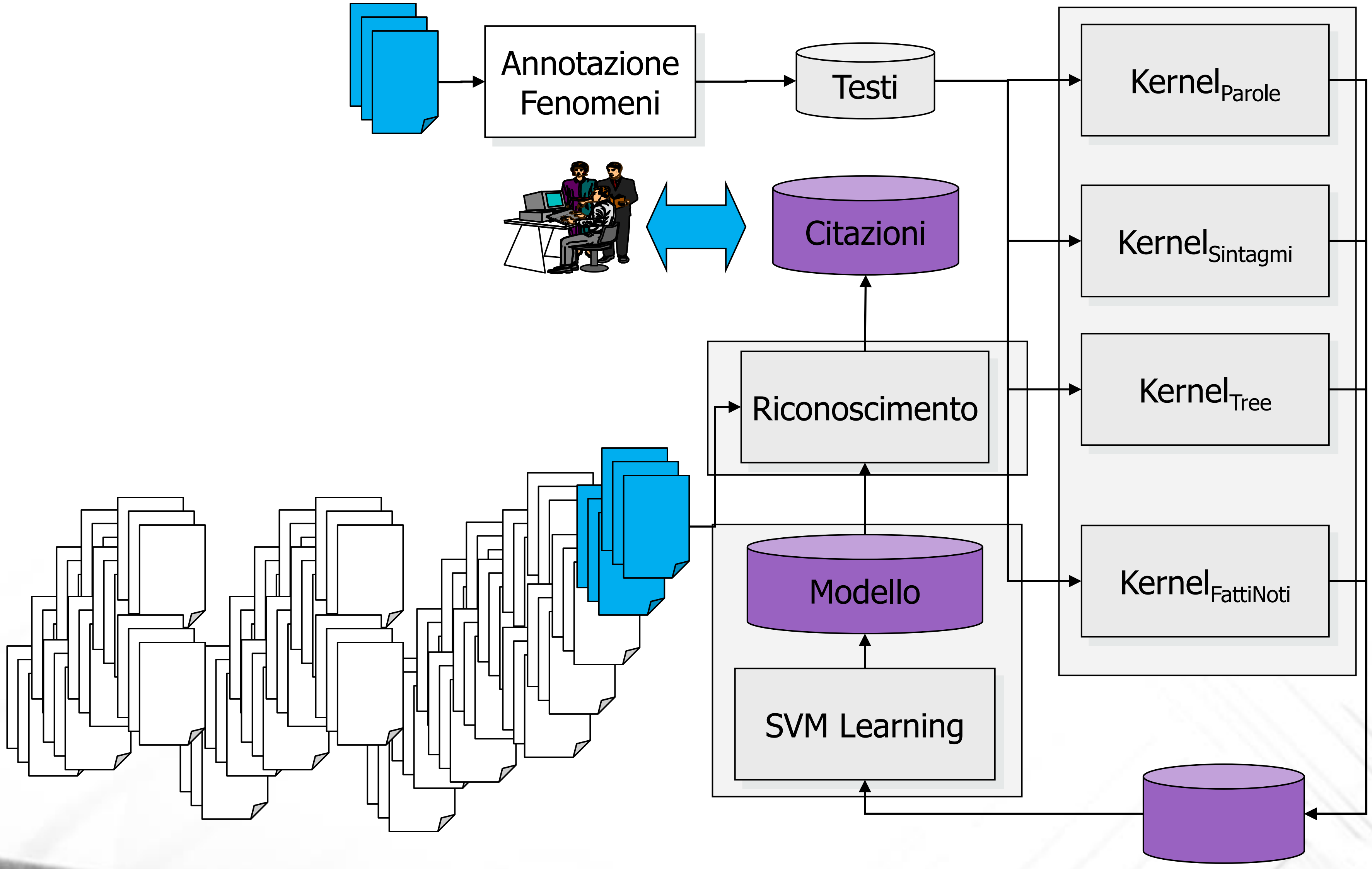
Il Processo Induttivo



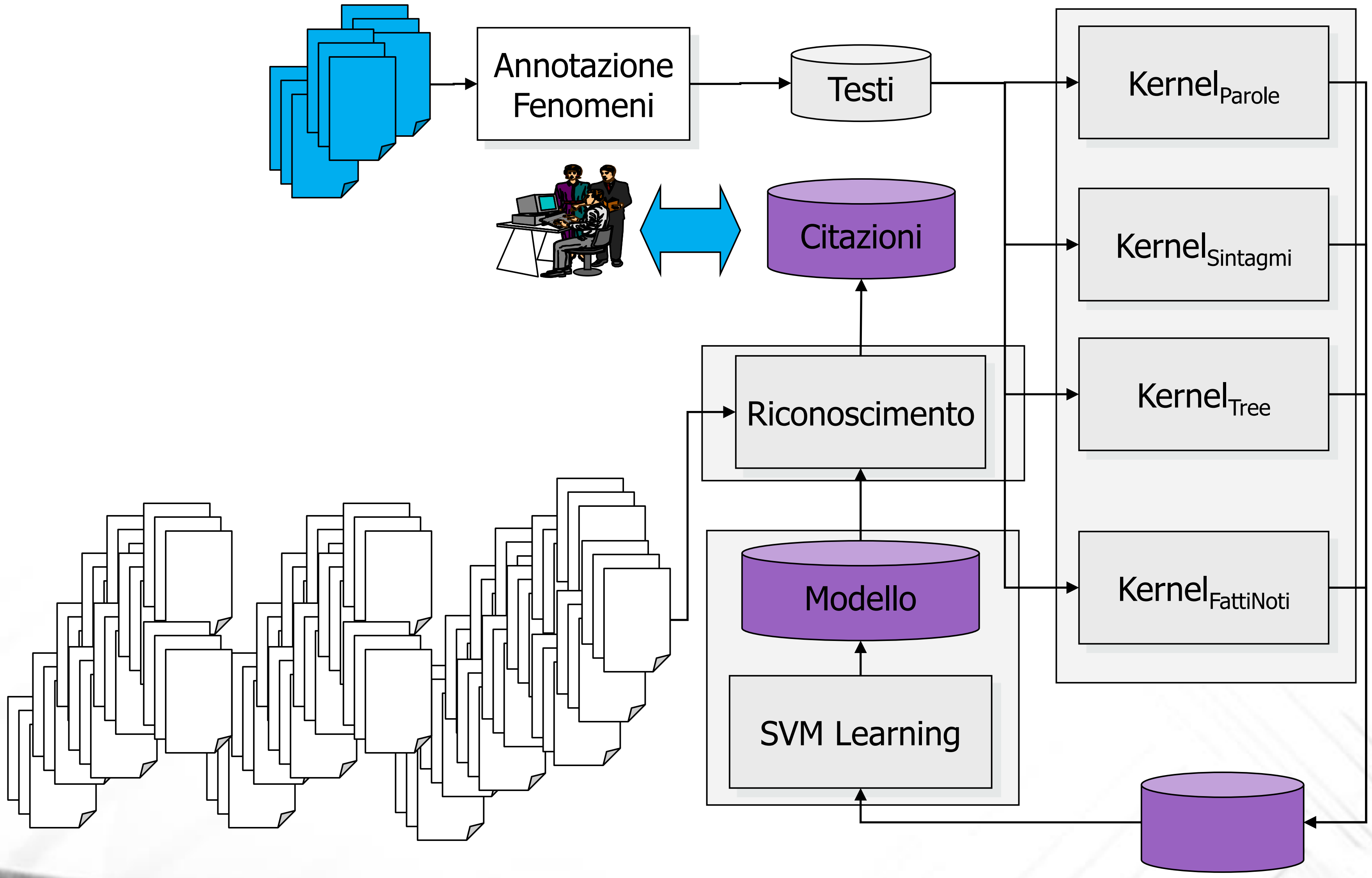
Supporto alla Analisi dei Dati



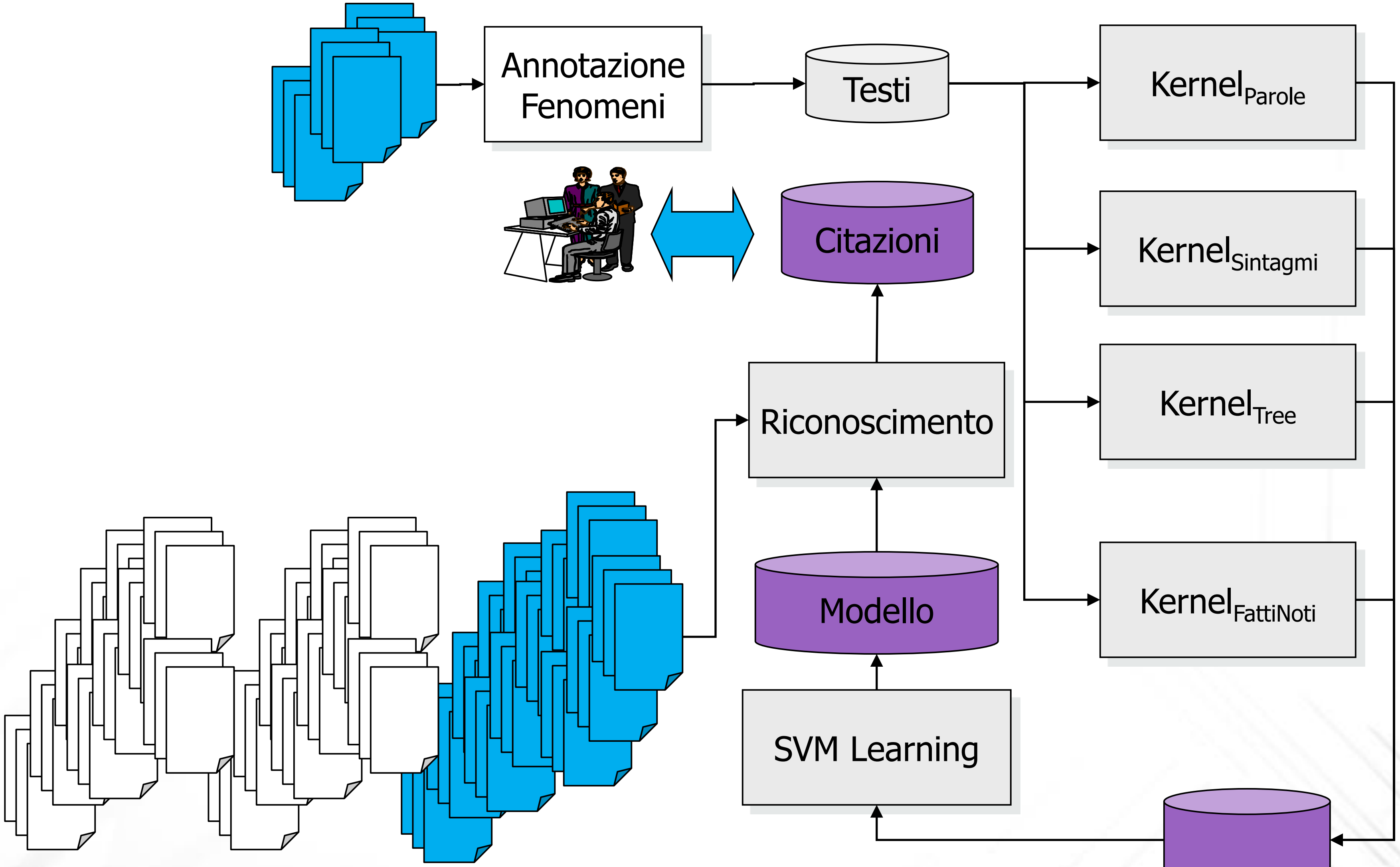
Supporto alla Analisi dei Dati



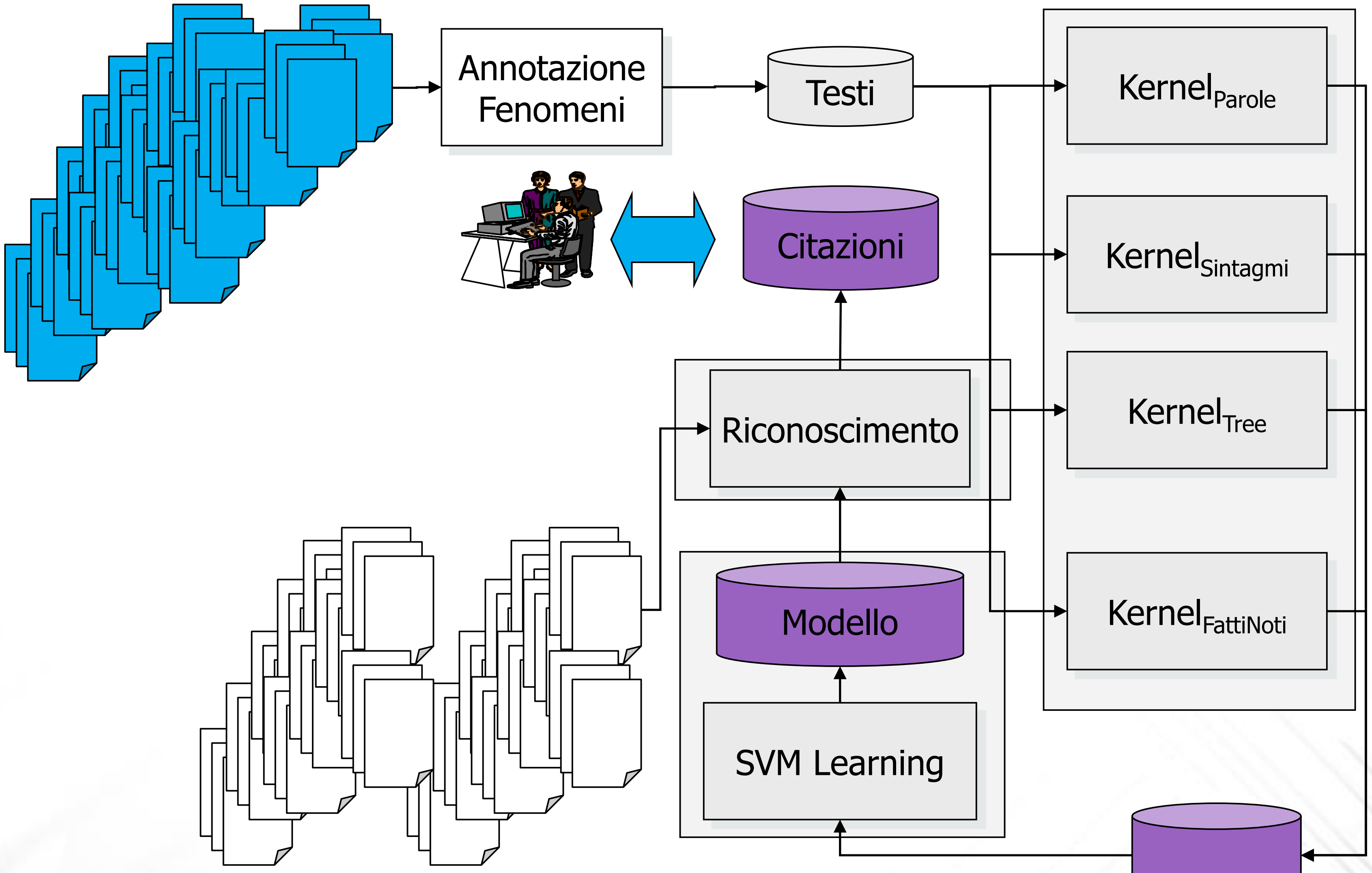
Supporto alla Analisi dei Dati



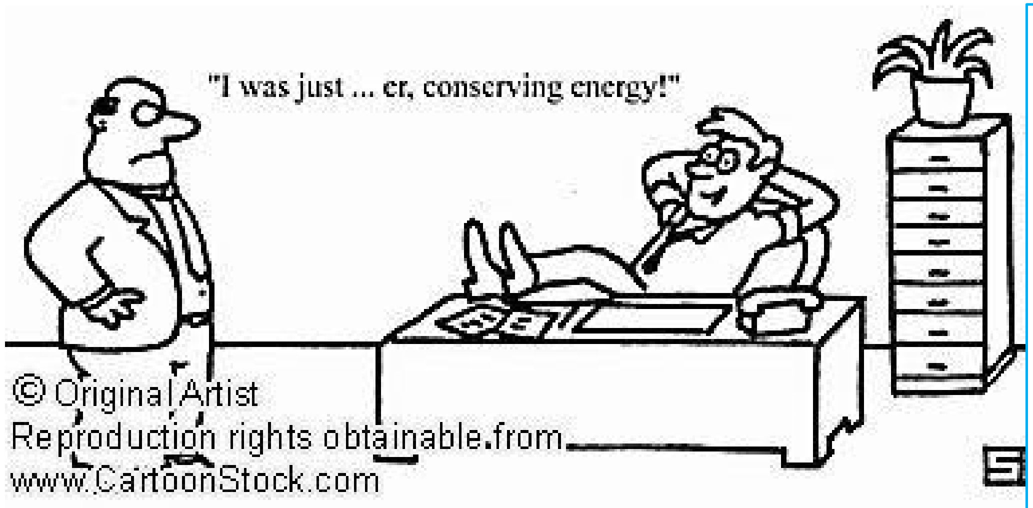
Supporto alla Analisi dei Dati



Supporto alla Analisi dei Dati



Tecnologie D*ata-driven* : Benefici



- Disponibilità di algoritmi molto **accurati** ed **efficienti**
- L'apprendimento è **portabile** mentre programmare i modelli è dipendente dal **task** (i.e. scenario)
- **Soluzioni ad alta qualità** possono essere ottenute in modo **cost-effective**
- **Raccogliere esempi** è più semplice e coinvolge **profili professionali meno specializzati**
- **L'analisi di larga scala** è resa possibile

Overview

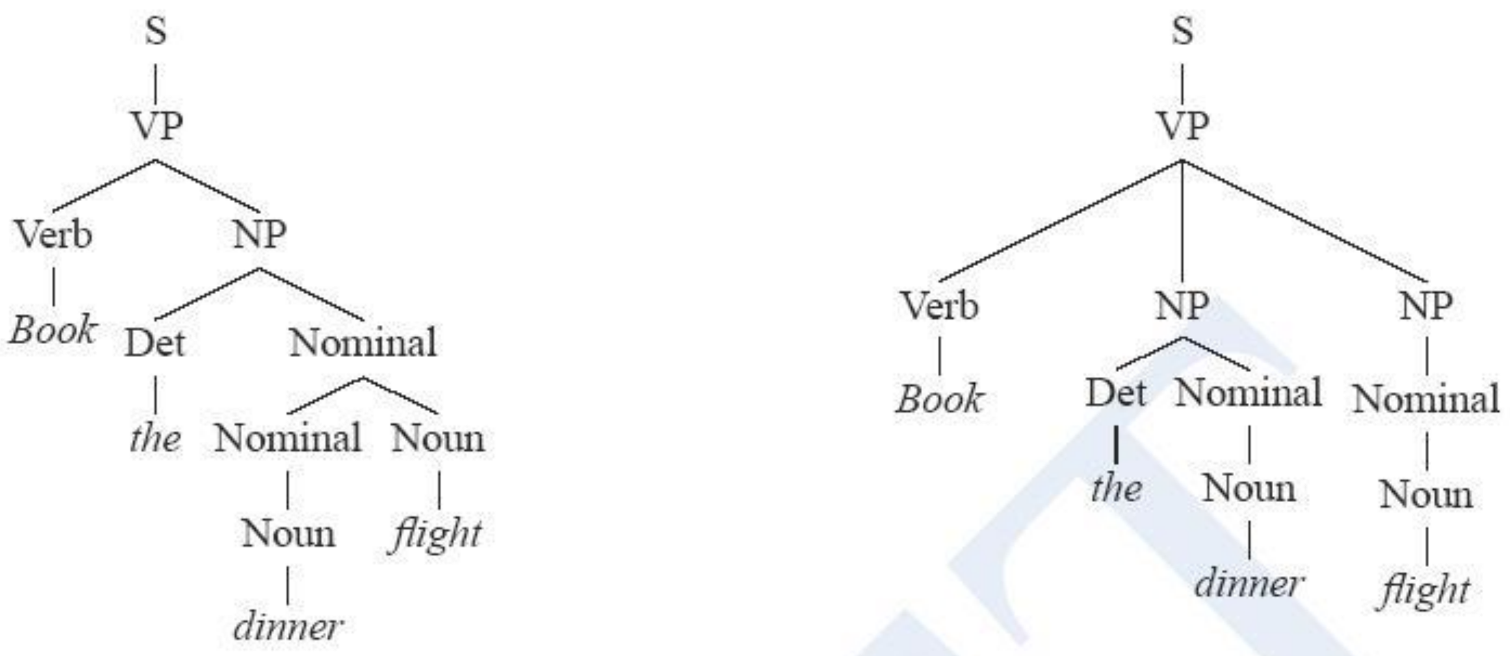
- Intelligenza Artificiale e Lingue parlate e scritte
  - Informazioni e Rappresentazioni coinvolte
  - Sfide (ri)correnti, battaglie (già) vinte e rischi inerenti …
- Elaborazione Automatica delle Lingue: Modelli, Metodi e *Risultati*
- *break*
- Ruolo delle Tecnologie dell'Apprendimento ed Applicazioni:
  - Sviluppo Automatico di Dizionari, Lessici Semantici ed Ontologie
  - Riconoscimento di fenomeni semantici
  - Trattamento Semantico della Documentazione Investigativa
  - Sistemi Web-based di Opinion Mining, Market Watch & Brand Reputation Management

Machine Learning in NLP

- Lexical Semantics:
  - Acquisition of lexical semantic dictionaries from corpora (aka distributional semantic methods, word spaces and embeddings)
  - Word Sense Disambiguation
- Data-driven Computational Semantics
  - Named Entity Recognition and Relation Extraction
  - Shallow Semantic Parsing (aka Semantic Role Labeling)
- NLP for Information Retrieval tasks
  - Semantic Indexing
  - (Open domain) Question Answering
  - Opinion Analysis
  - Community detection and Reccommeding

Le armi del Machine Learning

- Apprendimento di Regole e Pattern sui Dati
  - Frequent Pattern Mining (Basket analysis)
- Estensioni Probabilistiche delle Grammatiche
  - Probabilstic CFGs
  - Grammatiche Stocastiche
- Apprendimento Discriminativo nelle reti neurali
- SVM: percettroni
  - Funzioni Kernel in Spazi impliciti
- Modelli Bayesiani e Grafici



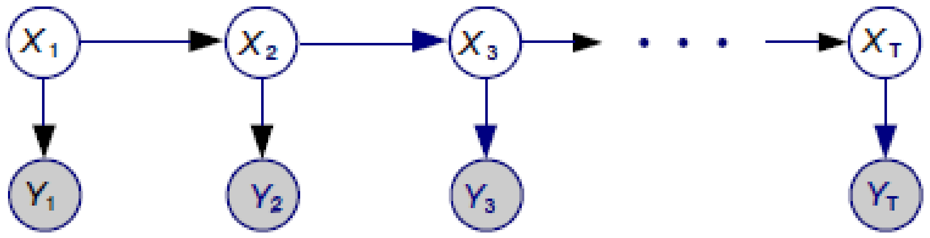
Rules			P
S	→	VP	.05
VP	→	Verb NP	.20
NP	→	Det Nominal	.20
Nominal	→	Nominal Noun	.20
Nominal	→	Noun	.75
Verb	→	book	.30
Det	→	the	.60
Noun	→	dinner	.10
Noun	→	flights	.40

Rules			P
S	→	VP	.05
VP	→	Verb NP NP	.10
NP	→	Det Nominal	.20
NP	→	Nominal	.15
Nominal	→	Noun	.75
Nominal	→	Noun	.75
Verb	→	book	.30
Det	→	the	.60
Noun	→	dinner	.10
Noun	→	flights	.40

Figure 13.2 Two parse trees for an ambiguous sentence. The transitive parse (a) cor-

Hidden Markov Models (HMM)

- Stati = Categorie/Concetti/Proprietà
- Osservazioni



- Emissioni

$$p(X_{1,\ldots,T}, Y_{1,\ldots,T}) = p(X_1)p(Y_1|X_1)\prod_{t=2}^{T}\left[p(X_t|X_{t-1})p(Y_t|X_t)\right]$$

- Transizioni

- Applicazioni:
  - Riconoscimento Vocale
  - Etichettatura Grammaticale (*POS tagging*)

Apprendimento Discriminativo

- Tipico delle reti neurali sin dalle prime proposte della Cibernetica (Minsky&Papert, 1956)
- Basato sulla nozione geometrica di prodotto interno e quindi di spazio vettoriale metrico
- Support Vector Machines (ma anche altri On-line Learning algorithms)
  - Kernels
  - Pre-training methods through word spaces and embeddings
  - Markovian SVMs for sequence labeling tasks
    - (SVM-HMM) come ibridazione di un modello discriminativo (SVM locali ai singoli time stamp) e di un approccio generativo (HMM per l'intera sequenza)

Named Entity Recognition

- See the Kozareva tutorial at: http://www.isi.edu/natural-language/teaching/cs544/spring11/kozareva_lecture3.ppt
- Continue ...

Semantic Role Labeling

- The Task
  - From Syntactic Argument Structures to Thematic Roles
  - SRL as a classification task
- SRL: Reference Linguistic Theories and Resources
- An SRL architecture
- Experiments and Results
  - Early models
  - SPTK (Croce et al., 2011)

Syntactic Argument Structures

- (Verbal) Relations require a fixed number of participants, called arguments
- The syntactic structure predicts the number and type of arguments through subcategorization frames
  - (Bob (gave (Mary) (the book) (on Monday)))
  - (Bob (gave (the book) (to Mary) (on Monday)))

Thematic roles

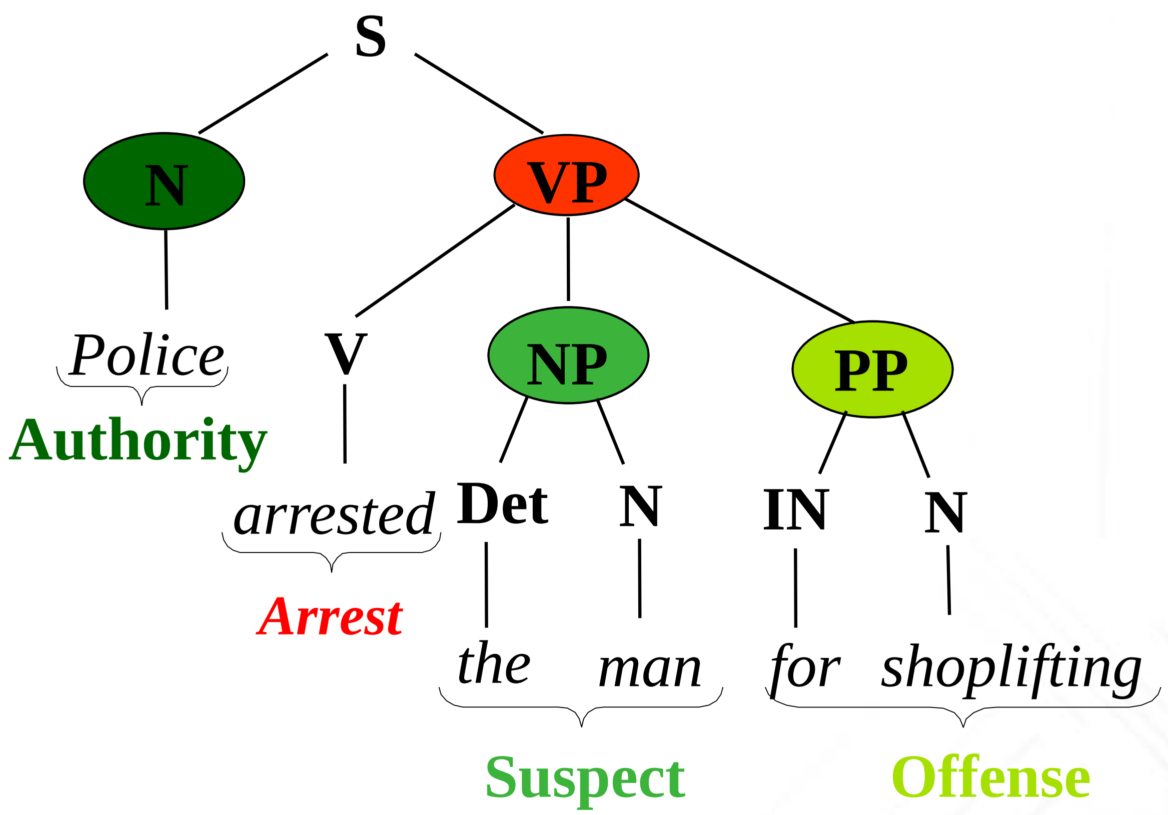
- Arguments play specific roles, called thematic roles, depending on the predicate but invariant across different syntactic structures giving rise to predicate argument structures
  - `give(Agent:` *Bob*, `Theme:` *the_book*, `Recipient:` *Mary*`)`

- Thematic roles of individual arguments <u>are indexed</u> by their predicates

- *General* and *lexicalized* roles have been introduced

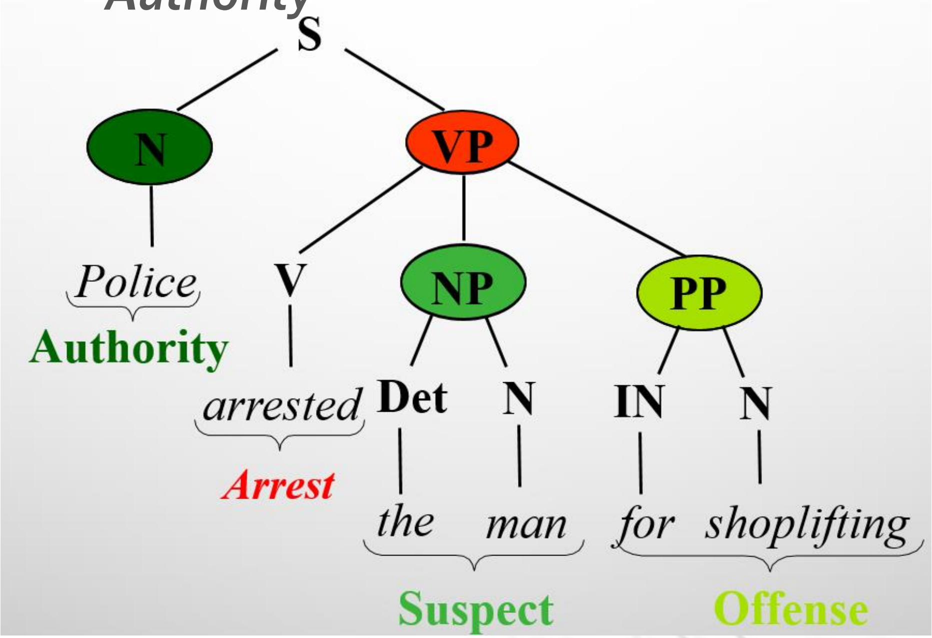
Linking syntax to semantics

- *Police arrested the man for shoplifting*



A tabular vision

- Word Predicate Semantic Role
- Police - Authority
- arrested Target Arrest
- the - SUSPECT
- man - SUSPECT
- for - OFFENSE
- Shoplifting - OFFENSE



Semantics in NLP: Resources

- Lexicalized Models
  - Propbank
  - NomBank
- Framenet
  - Inspired by frame semantics
  - Frames are lexicalized prototoypes for real -world situations
  - Participants are called frame elements (roles)

Frame Semantics

- Research in Empirical Semantics suggests that words represents categories of experience (*situations*)
- A *frame* is a cognitive structuring device (i.e. a kind of prototype) indexed by *words* and used to support understanding (Fillmore, 1975)
  - *Lexical Units* evoke a Frame in a sentence
- Frames are made of *elements* that express participants to the situation (*Frame Elements*)
- During communication LUs evoke the frames

Frame Semantics

Frame: KILLING		
A KILLER or CAUSE causes the death of the VICTIM.		
Frame Elements	KILLER	**John** drowned Martha.
	VICTIM	John drowned **Martha**.
	MEANS	The flood exterminated the rats **by cutting off access to food**.
	CAUSE	**The rockslide** killed nearly half of the climbers.
	INSTRUMENT	It's difficult to suicide **with only a pocketknife**.
Predicates	annihilate.v, annihilation.n, asphyxiate.v,assassin.n, assassinate.v, assassination.n, behead.v, beheading.n, blood-bath.n, butcher.v, butchery.n, carnage.n, crucifixion.n, crucify.v, deadly.a, decapitate.v, decapitation.n, destroy.v, dispatch.v, drown.v, eliminate.v, euthanasia.n, euthanize.v, …	

The FrameNet project

- The aims
  - Create a lexical resource by describing a significant portion of English in terms of precise and rich frame semantics
- The output
  - Frame Database: a structured system of Frames and Fes
  - Lexical database: syntactic and semantic descriptions of frame-evoking words (N,V,A)
  - Annotated Corpus: wide coverage examples

Killing

D

A

F

C

C

E

Instru

Seman

Exclu

Killer

Exclu

Mean

Seman

Exclu

Victin

Seman

FEs:

Non-Core:

Beneficiary [ben]	This extra-thematic FE applies to participants that derive a benefit from the occurrence of the event specified by the target predicate.
Circumstances []	Circumstances describe the state of the world (at a particular time and place) which is specifically independent of the event itself and any of its participants.

Semantic Type: Physical_entity — It's difficult to SUICIDE with only a pocketknife.

Excludes: Cause

Killer [Kill] — The person or sentient entity that causes the death of the Victim.

Excludes: Cause

Means [] — The method or action that the Killer or Cause performs resulting in the death of the Victim.

Semantic Type: State_of_affairs — The flood EXTERMINATED the rats by cutting off access to food.

Excludes: Cause

Victim [] — The living entity that dies as a result of the killing.

Semantic Type: Sentient

Non-Core:

Beneficiary [ben] — This extra-thematic FE applies to participants that derive a benefit from the occurrence of the event specified by the target predicate.

Framenet - Data

- Methodology of constructing FrameNet
  - Define/discover/describe frames
  - Decide the participants (frame elements)
  - List lexical units that evoke the frame
  - Find example sentences in the BNC and annotate them
- Corpora
  - FrameNet I -British National Corpus only
  - FrameNet II -LDC North American Newswire corpora
- Size
  - >10,000 lexical units, >825 frames, >135,000 sentences
- http://framenet.icsi.berkeley.edu

Frame Data & Domains (from G&J,2002)

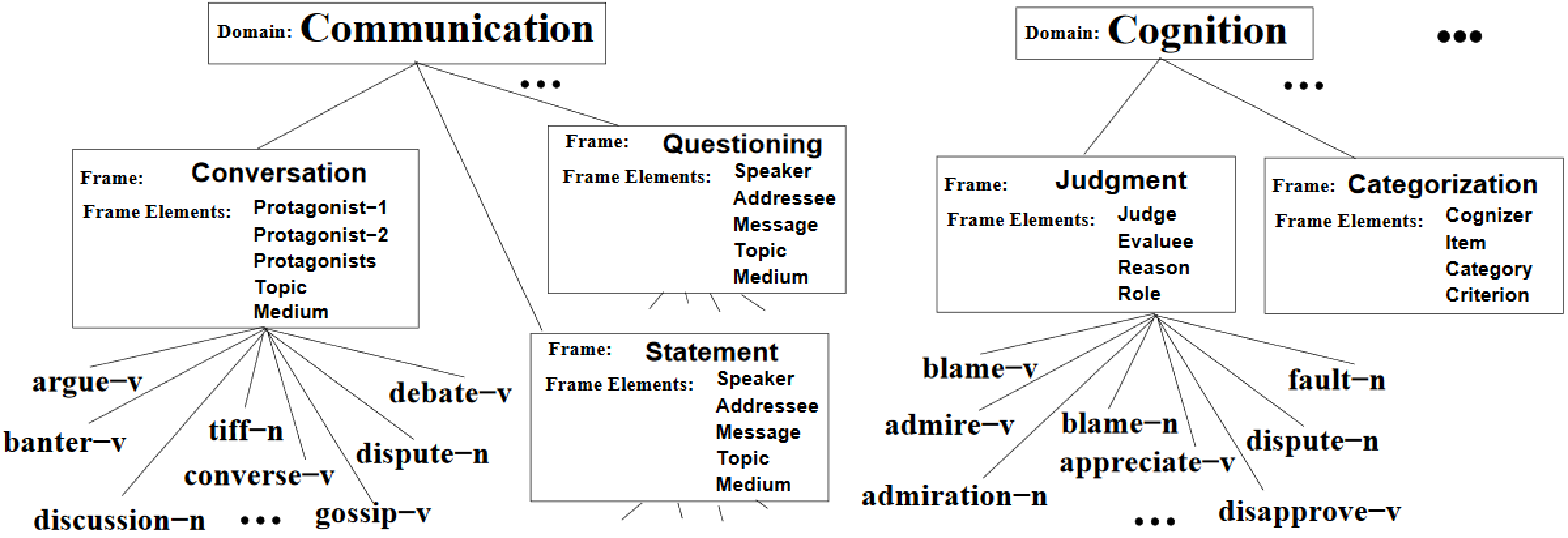
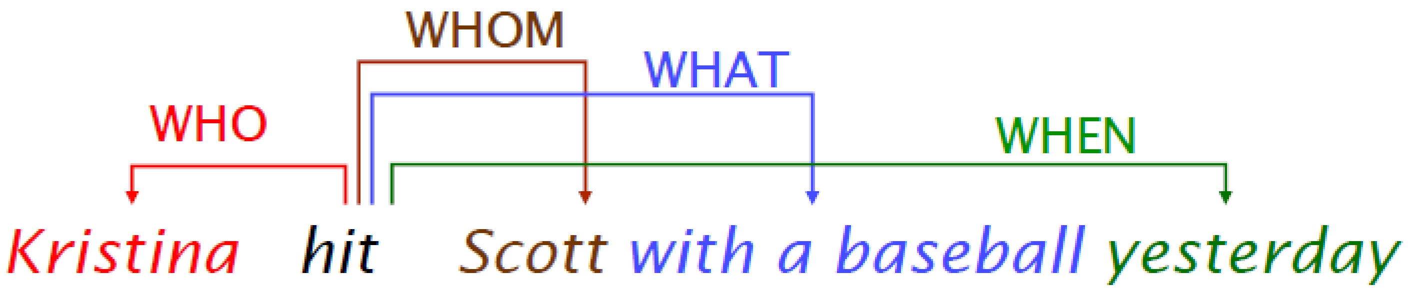


Figure 1
Sample domains and frames from the FrameNet lexicon.

Recognizing Predicates: SRL

- Semantic role labeling vs. QA



- Who hit Scott with a baseball?
- Whom did Kristina hit with a baseball?
- What did Kristina hit Scott with?
- When did Kristina hit Scott with a baseball?

Roles and variants in QA

Yesterday, Kristina hit Scott with a baseball

Scott was hit by Kristina yesterday with a baseball

Yesterday, Scott was hit with a baseball by Kristina

With a baseball, Kristina hit Scott yesterday

Yesterday Scott was hit by Kristina with a baseball

Kristina hit Scott with a baseball yesterday

- Agent, hitter
- Thing hit
- Instrument
- Temporal adjunct

SRL: task formulation

- Most general formulation: determine a labeling on (usually but not always contiguous) *substrings* (*phrases*) of the sentence *s*, given a predicate *p*

[$_{A0}$ The queen] **broke** [$_{A1}$ the window].

[$_{A1}$ By working hard], [$_{A0}$ he] **said**, [$_{C\text{-}A1}$ you can get exhausted].

- Every substring *c* can be represented by a set of word indices $c \subseteq \{1, 2, \ldots, m\}$
- More formally, a semantic role labeling is a mapping from the set of substrings of s to the label set ***L***. ***L*** includes all argument labels and NONE.

The SRL cascade

- Identification:
  - Very hard task: to separate the argument substrings from the rest in this exponentially sized set
  - Usually only 1 to 9 (avg. **2.7**) substrings have labels ARG and the rest have NONE for a predicate
- Classification:
  - Given the set of substrings that have an *ARG* label, decide the exact semantic label
- Core argument semantic role labeling: (easier)
  - Label phrases with core argument labels only. The modifier arguments are assumed to have label NONE.

ML Approaches

- **Local models** decide the label of each substring independently of the labels of other substrings

- This can lead to inconsistencies
  - overlapping argument strings

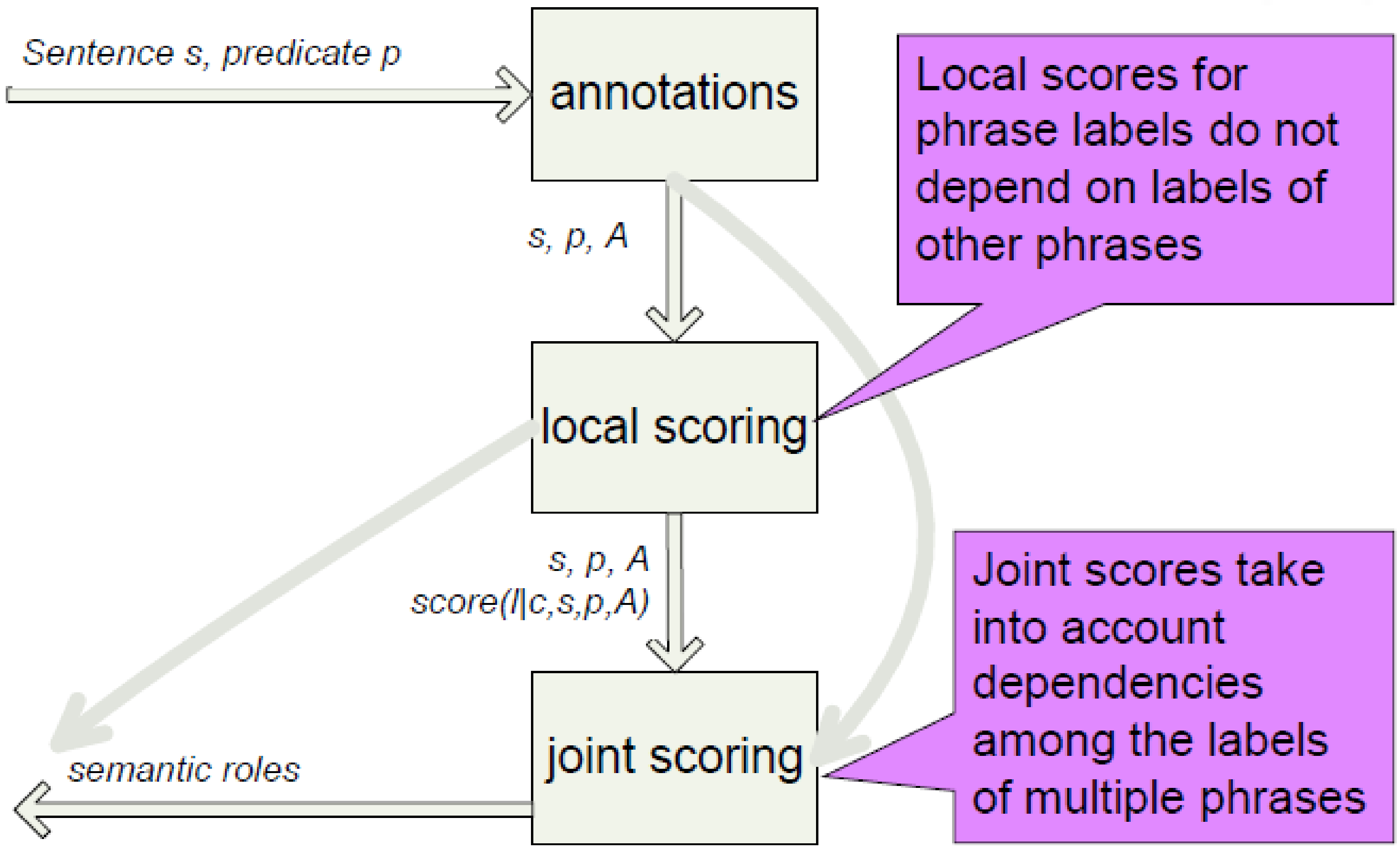
    By [$_{A1}$ working [$_{A1}$ hard] , he] ***said*** *, you can achieve a lot.*
  - repeated arguments

    By [$_{A1}$ working] hard , [$_{A1}$ he] ***said*** *, you can achieve a lot.*
  - missing arguments

    [$_{A0}$ By working hard , he] **said** , [$_{A0}$ you can achieve a lot].

- **Joint models** take into account the dependencies among labels of different substrings

The general SRL architecture



Previous work on Local …

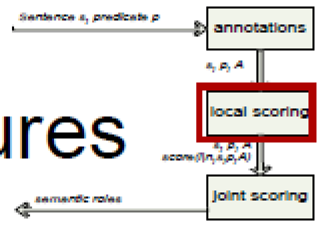
- [Gildea&Jurafsky 02]
  - **Identification + Classification** for local scoring experiments
  - **One Step** for joint scoring experiments
- [Xue&Palmer 04] and [Punyakanok et al. 04, 05]
  - **Pruning + Identification + Classification**
- [Pradhan et al. 04] and [Toutanova et al. 05]
  - **One Step**

… and Joint SRL models

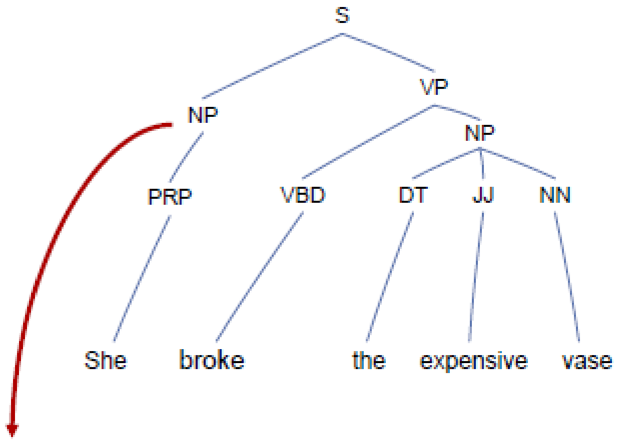
- Tight integration of local and joint scoring in a single probabilistic model and exact search [Cohn&Blunsom 05] [Màrquez et al. 05],[Thompson et al. 03]
  - When the joint model makes strong independence assumptions
- Re-ranking or approximate search to find the labeling which maximizes a combination of local and a joint score [Gildea&Jurafsky 02] [Pradhan et al. 04] [Toutanova et al. 05] [Moschitti et al. 07]
  - Usually exponential search required to find the exact maximizer
- Exact search for best assignment by local model satisfying hard joint constraints
  - Using Integer Linear Programming [Punyakanok et al 04,05] (worst case NP-hard)

Features (for Local models)

Gildea & Jurafsky (2002) Features



- Key early work
  - Future systems use these features as a baseline
- Constituent Independent
  - Target predicate (lemma)
  - Voice
  - Subcategorization
- Constituent Specific
  - Path
  - Position (*left, right*)
  - Phrase Type
  - Governing Category (*S* or *VP*)
  - Head Word



Target	*broke*
Voice	*active*
Subcategorization	*VP→VBD NP*
Path	*VBD↑VP↑S↓NP*
Position	*left*
Phrase Type	*NP*
Gov Cat	*S*
Head Word	*She*

Application of distributional lexicons for Semantic Role Labeling @ UTV

- An important application of SVM is Semantic Role labeling wrt Propbank or Framenet
- In the UTV system, a cascade of classification steps is applied:
  - Predicate detection
  - Boundary recognition (Argument Identification)
  - Argument categorization (Local models)
  - Reranking (Joint model)
- Input: a sentence and its parse trees

Tree kernels for SRL

- See «Short Introduction to Semantic Tree Kernels»

Semantic Role Labeling via SVM Learning

- Three steps:
  - Predicate Detection:
    - Locate occurrences of frames in sentences
    - Recognition of predicate words or multiword expressions
  - Boundary Detection
    - One binary classifier applied to the parse tree nodes
  - Argument Type Classification
    - Multi-classification problem, where n binary classifiers are applied, one for each argument class (i.e. frame element)
    - They are combined in a ONE-vs-ALL scheme, i.e. the argument type that is categorized by an SVM with the maximum score is selected

Automatic Predicate Argument Extraction

Deriving Positive/Negative examples

- Given a sentence, a predicate *p*:
  - Derive the sentence parse tree
  - For each node pair $<N_p,N_x>$
    - Extract a feature representation set *F*
    - If N_x exactly covers the i-th argument, *Arg-i*, *F* is one of the positive examples for an *Arg-I* classifier
    - *F* is a negative example for *Arg-I,* otherwise

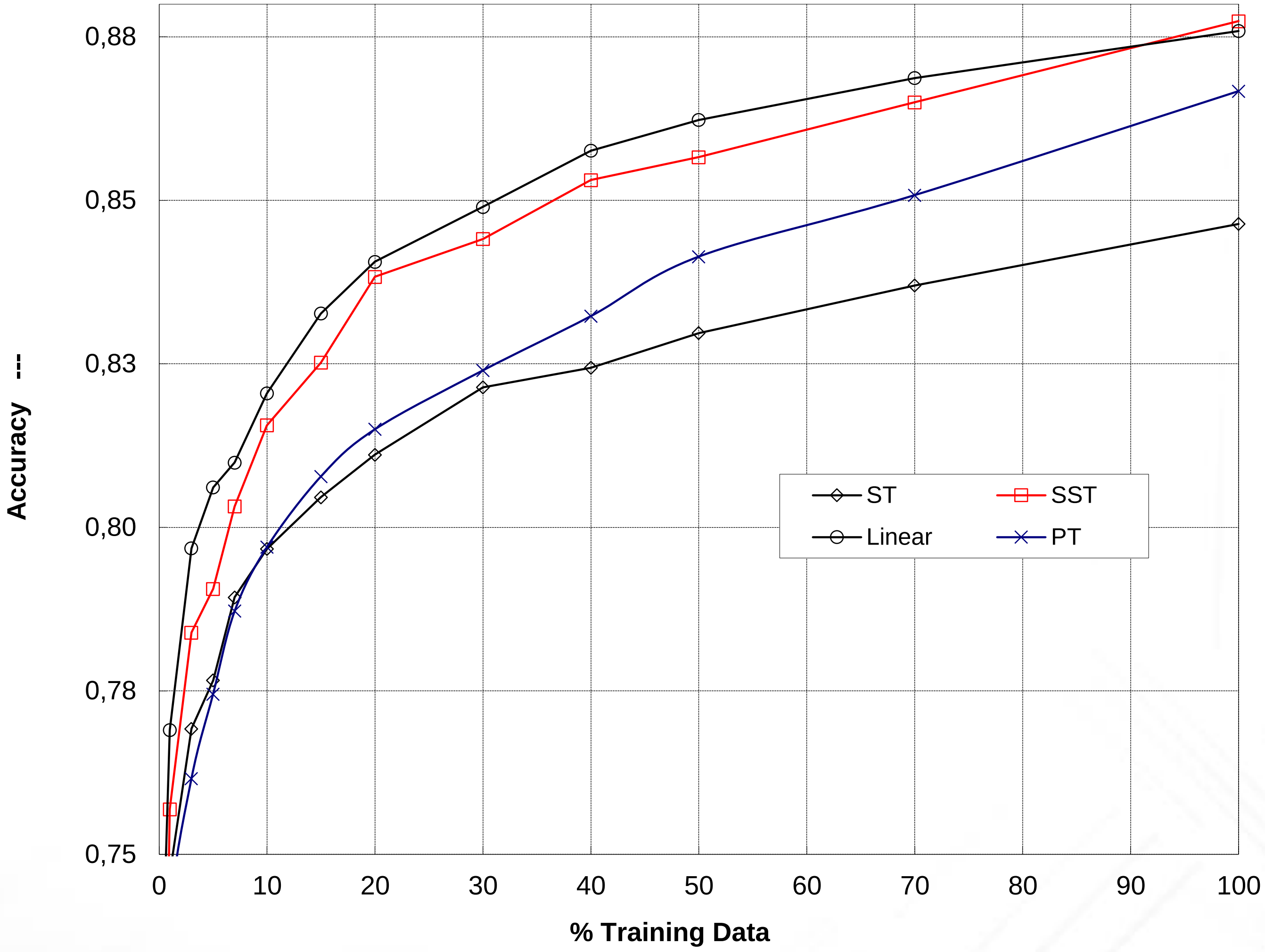
SRL at RTV: Smoothed Partial Tree Kernels

- Experimental Set-up (Croce et al., EMNLP 2011)

- FrameNet version: 1.3
- 271,560 training and 30,173 test examples respectively
- LTH dependency parser (Malt, Johansson & Nugues, 2007).
- Word space: LSA applied to the BNC corpus (about 10M words).

- Number of targeted frames: 648 frames
- Parse trees format: GRCT and LCT
- A total of 4,254 binary role classifiers (RC)

Argument Classification Accuracy



SRL in Framenet: Results

Eval Setting	Tree Kernels P	Tree Kernels R	Tree Kernels F_1	Tree Kernels + PK P	Tree Kernels + PK R	Tree Kernels + PK F_1
				PK alone		
BD	-	-	-	.887	.675	.767
BD Proj.	-	-	-	.850	.647	.735
BD+RC	-	-	-	.654	.498	.565
BD+RC Proj.	-	-	-	.625	.476	.540
	TK			**TK + PK**		
BD	.949	.652	.773	.915	.698	.792
BD Proj.	.919	.631	.748	.875	.668	.758
BD+RC	.697	.479	.568	.680	.519	.588
BD+RC Proj.	.672	.462	.548	.648	.495	.561
	TKL			**TKL + PK**		
BD	.938	.659	.774	.908	.701	.791
BD Proj.	.906	.636	.747	.868	.670	.757
BD+RC	.689	.484	.569	.675	.521	.588
BD+RC Proj.	.663	.466	.547	.644	.497	.561

Table 4.1: Results on FrameNet dataset. The table shows Precision, Recall, and F-measure achieved by the Polynomial Kernel (PK) and two different Tree Kernels (TK and TKL). Also, results for their combinations are shown. All experiments exploit 2% training data for Boundary Detection, and 90% for Role Classification.

SRL in Framenet: Results

Eval Setting	Tree Kernels P	Tree Kernels R	Tree Kernels F_1	Tree Kernels + PK P	Tree Kernels + PK R	Tree Kernels + PK F_1
				PK alone		
BD	-	-	-	.887	.675	.767
BD Proj.	-	-	-	.850	.647	.735
BD+RC	-	-	-	.654	.498	.565
BD+RC Proj.	-	-	-	.625	.476	.540
	TK			**TK + PK**		
BD	.949	.652	.773	.915	.698	.792
BD Proj.	.919	.631	.748	.875	.668	.758
BD+RC	.697	.479	.568	.680	.519	.588
BD+RC Proj.	.672	.462	.548	.648	.495	.561
	TKL			**TKL + PK**		
BD	.938	.659	.774	.908	.701	.791
BD Proj.	.906	.636	.747	.868	.670	.757
BD+RC	.689	.484	.569	.675	.521	.588
BD+RC Proj.	.663	.466	.547	.644	.497	.561

Table 4.1: Results on FrameNet dataset. The table shows Precision, Recall, and F-measure achieved by the Polynomial Kernel (PK) and two different Tree Kernels (TK and TKL). Also, results for their combinations are shown. All experiments exploit 2% training data for Boundary Detection, and 90% for Role Classification.

Framenet SRL: best results

- Best system [Erk&Pado, 2006]
  - 0.855 Precision, 0.669 Recall
  - 0.751 F1
- Trento (+RTV) system (Coppola, PhD2009)

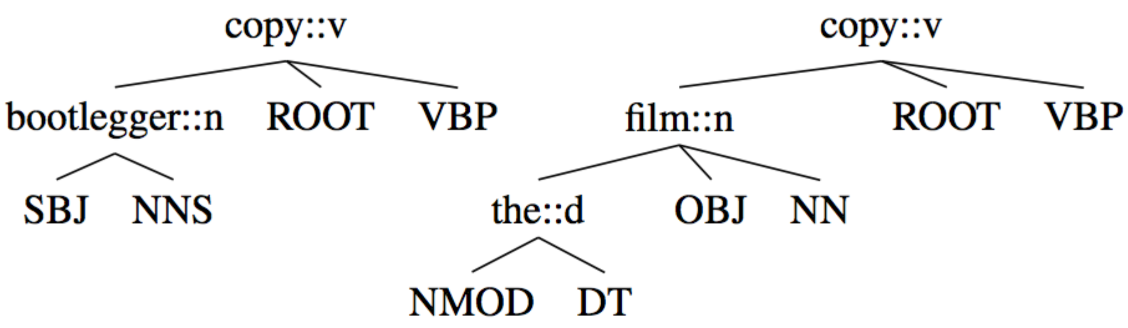
Enhanced PK+TK			
Eval Setting	P	R	F_1
BD (nodes)	1.0	.732	.847
BD (words)	.963	.702	.813
BD+RC (nodes)	.784	.571	.661
BD+RC (words)	.747	.545	.630

Table 4.2: Results on the FrameNet dataset. Best configuration from Table 4.1, raised to 90% of training data for BD and RC.

Argument Classification (Croce et al., 2013)

- UTV experimented with a FrameNet SRL classification (gold standard boundaries)
- We used the FrameNet version 1.3: 648 frames are considered
  - Training set: 271,560 arguments (90%)
  - Test set: 30,173 arguments (10%)

[*Bootleggers*]$_{CREATOR}$, *then* **copy** [*the film*]$_{ORIGINAL}$ [*onto hundreds of VHS tapes*]$_{GOAL}$



Kernel	Accuracy
GRCT	87,60%
$GRCT_{LSA}$	88,61%
LCT	87,61%
LCT_{LSA}	88,74%
GRCT+LCT	87,99%
$GRCT_{LSA}$+LCT_{LSA}	**88,91%**

Overview

- Intelligenza Artificiale e Lingue parlate e scritte
  - Informazioni e Rappresentazioni coinvolte
  - Sfide (ri)correnti, battaglie (già) vinte e rischi inerenti …
- Elaborazione Automatica delle Lingue: Modelli, Metodi e *Risultati*
- *break*
- Ruolo delle Tecnologie dell'Apprendimento ed Applicazioni:
  - Sviluppo Automatico di Dizionari, Lessici Semantici ed Ontologie
  - Riconoscimento di fenomeni semantici
  - Trattamento Semantico della Documentazione Investigativa
  - Sistemi Web-based di Opinion Mining, Market Watch & Brand Reputation Management

Distributional Semantics: the overall process

Corpus Normalization

- Remove Irrelevant Information (e.g. mark-ups)
- **Segment Texts into coherent units**

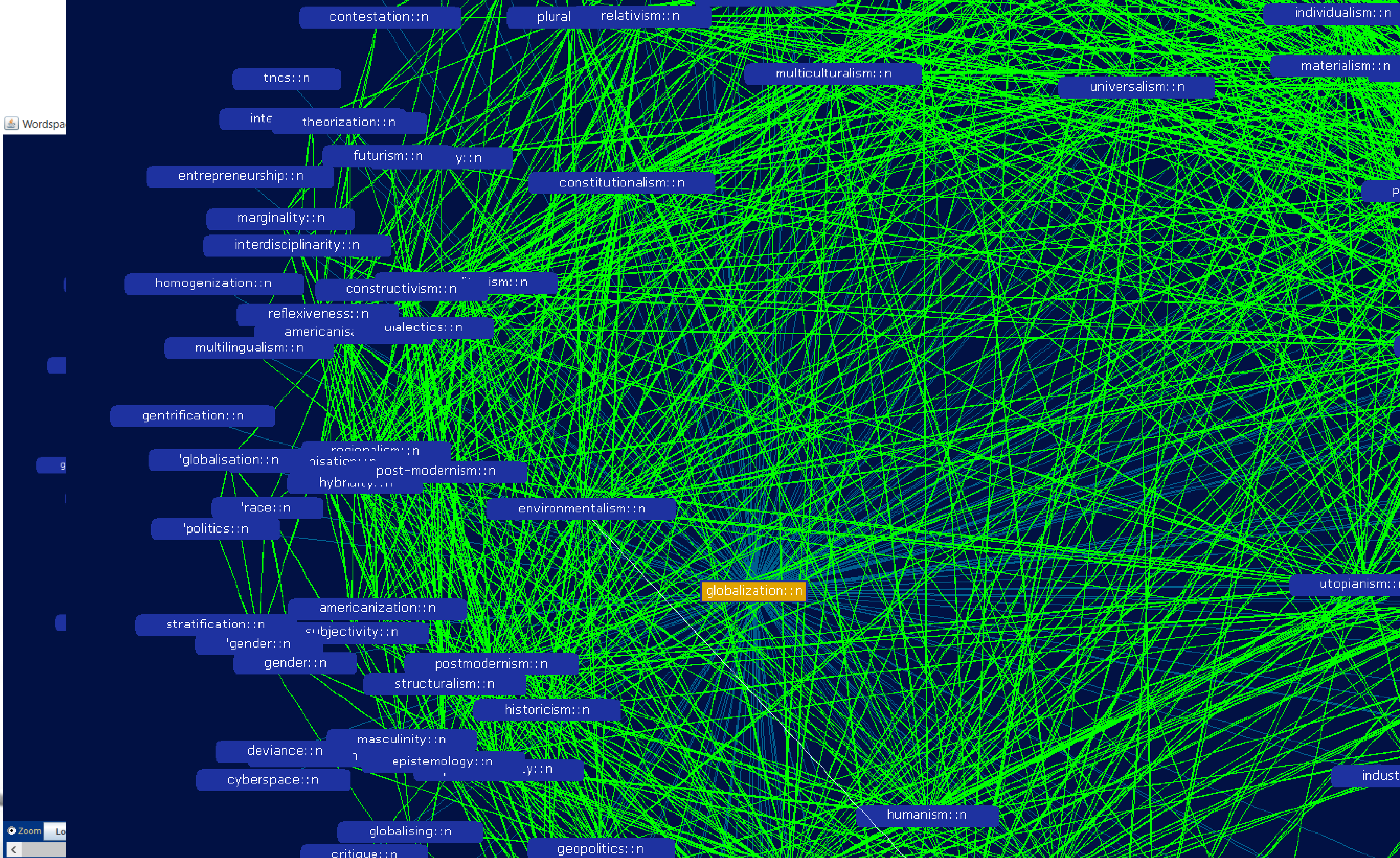
Lemmatization & Counting

- Apply **Shallow Linguistic Processing** (POS tagging)
- Build the co-occurrence **word-by-context matrix**

PCA Analysis

- Apply **SVD decomposition** of order *k*
- Map individual lexical entries into k-dimensional real-valued vectors

Latent Semantic Analysis (LSA), (Landauer & Dumais, 1997)



contestation::n
plural
relativism::n
individualism::n
tncs::n
multiculturalism::n
materialism::n
universalism::n
theorization::n
futurism::n
entrepreneurship::n
constitutionalism::n
marginality::n
interdisciplinarity::n
homogenization::n
constructivism::n
reflexiveness::n
multilingualism::n
gentrification::n
'globalisation::n
post-modernism::n
environmentalism::n
'race::n
'politics::n
globalization::n
utopianism::n
americanization::n
stratification::n
'gender::n
gender::n
postmodernism::n
structuralism::n
historicism::n
masculinity::n
deviance::n
epistemology::n
cyberspace::n
humanism::n
globalising::n
geopolitics::n
Zoom


he is running the Marathon this year
play::v$$0.505
host::v$$0.550
build::v$$0.574
call::v$$0.523
visit::v$$0.519
attend::v$$0.521
close::v$$0.520
base::v$$0.495
hold::v$$0.621
support::v$$0.527
open::v$$0.632
She is running a relief operation in Sudan
enter::v$$0.568
deliver::v$$0.500
run::v
return::v$$0.506
reach::v$$0.563
offer::v$$0.496
move::v$$0.626
leave::v$$0.505
install::v$$0.520
turn::v$$0.608
set::v$$0.605
take::v$$0.585
go::v$$0.581
connect::v$$0.510
start::v$$0.645
put::v$$0.530
Running a new program on a PC
get::v$$0.495
The children ran to the store

Automatic Acquisition of distributional semantic lexicons from corpora

- Three main approaches
  - Bayesian models, e.g. Topic models or LDA
  - Algebraic models, ususally based on matrix decomposition (e.g. LSA)
  - Neural models, e.g. self-associative (auto)encoders (Mitkov, 2013)

- All methods output n-dimensional lexical vectors that corresponds to units of semantic descriptions
- The overall vector set is called *word embedding* and it corresponds to an implicit representation of the mental lexicon

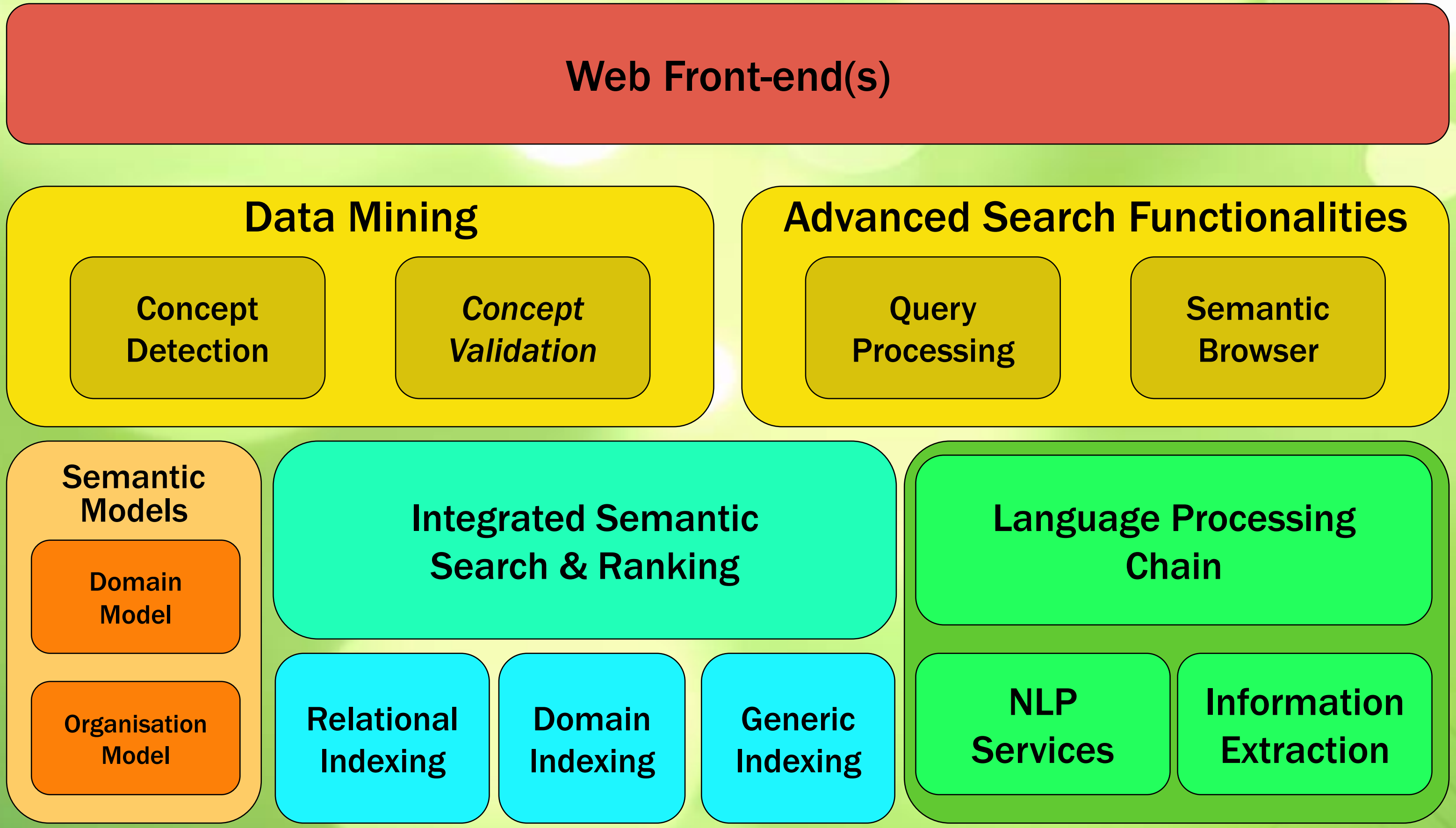
Semantics, Natural Language & Learning

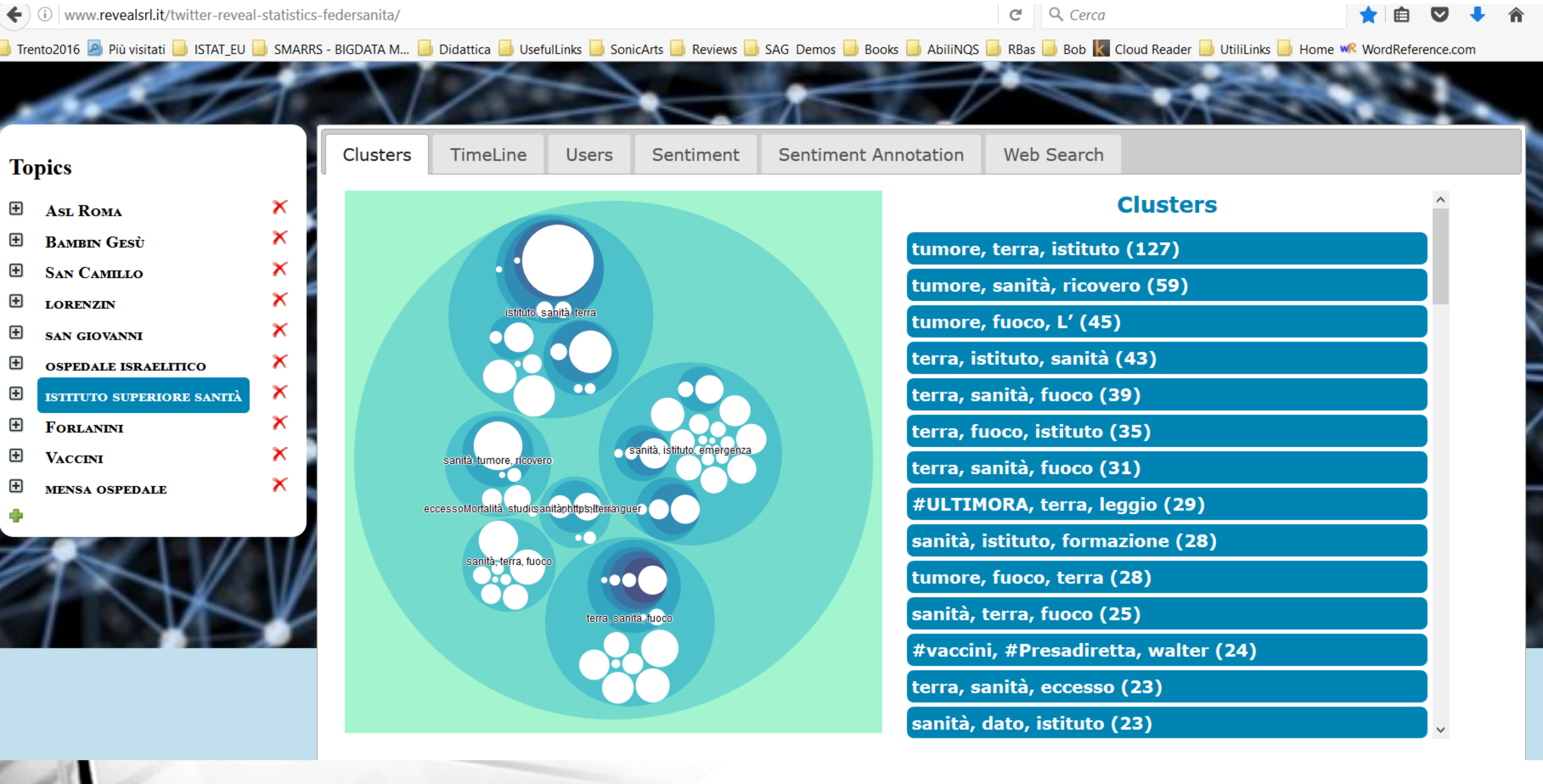
- From **Learning to Read** to **Knowledge Distillation** as a (integrated pool of) Semantic interpretation Task(s)
  - **Information Extraction**
    - Entity Recognition and Classification
    - Relation Extraction
    - Semantic Role Labeling (Shallow Semantic Parsing)
  - **Estimation of Text Similarity**
    - Structured Text Similarity/Textual Entailment Recognition
    - Sense disambiguation
  - **Semantic Search, Question Classification and Answer Ranking**
  - **Knowledge Acquisition**, e.g. ontology learning
  - **Social Network Analysis, Opinion Mining**

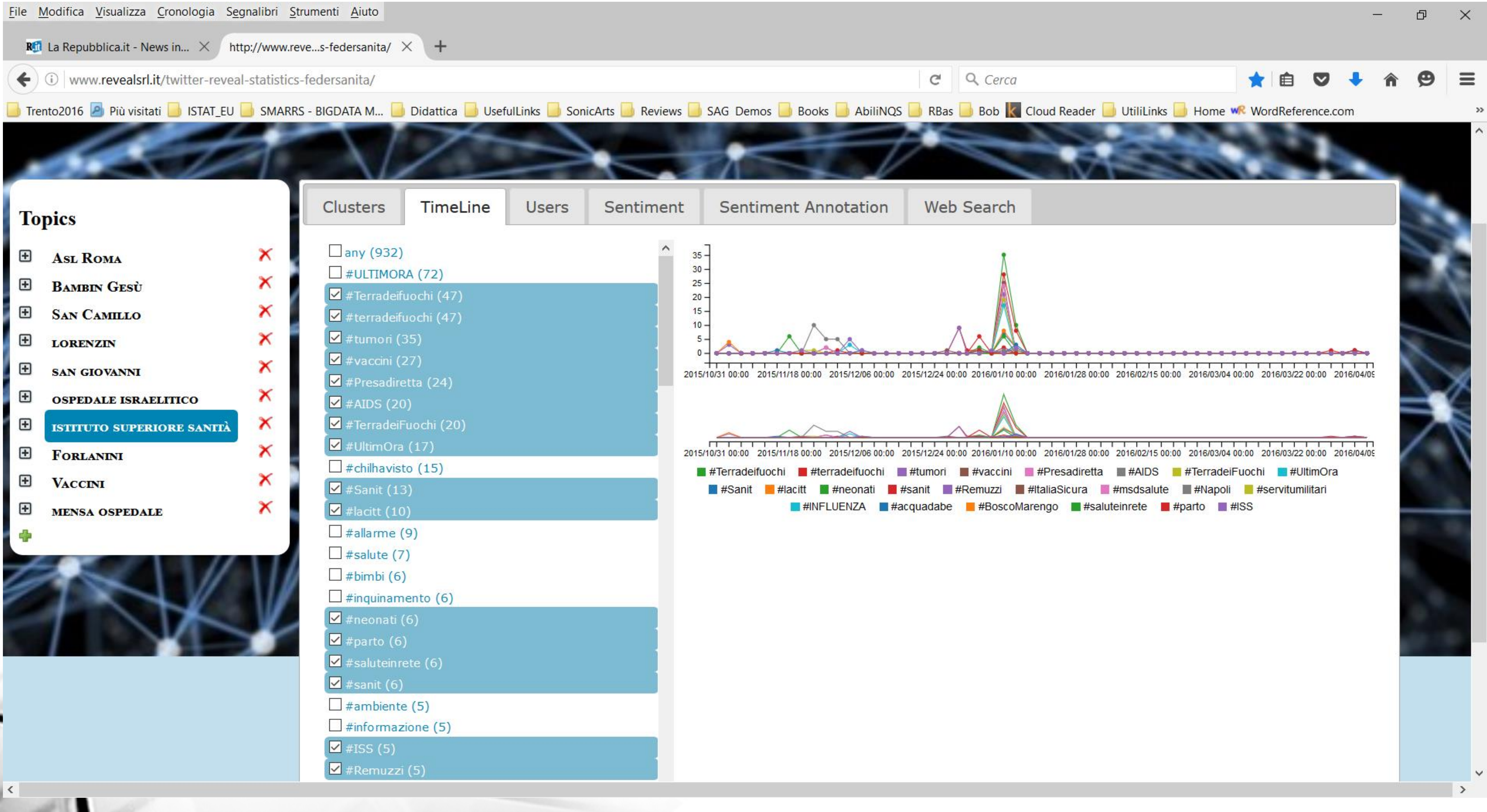
Overview

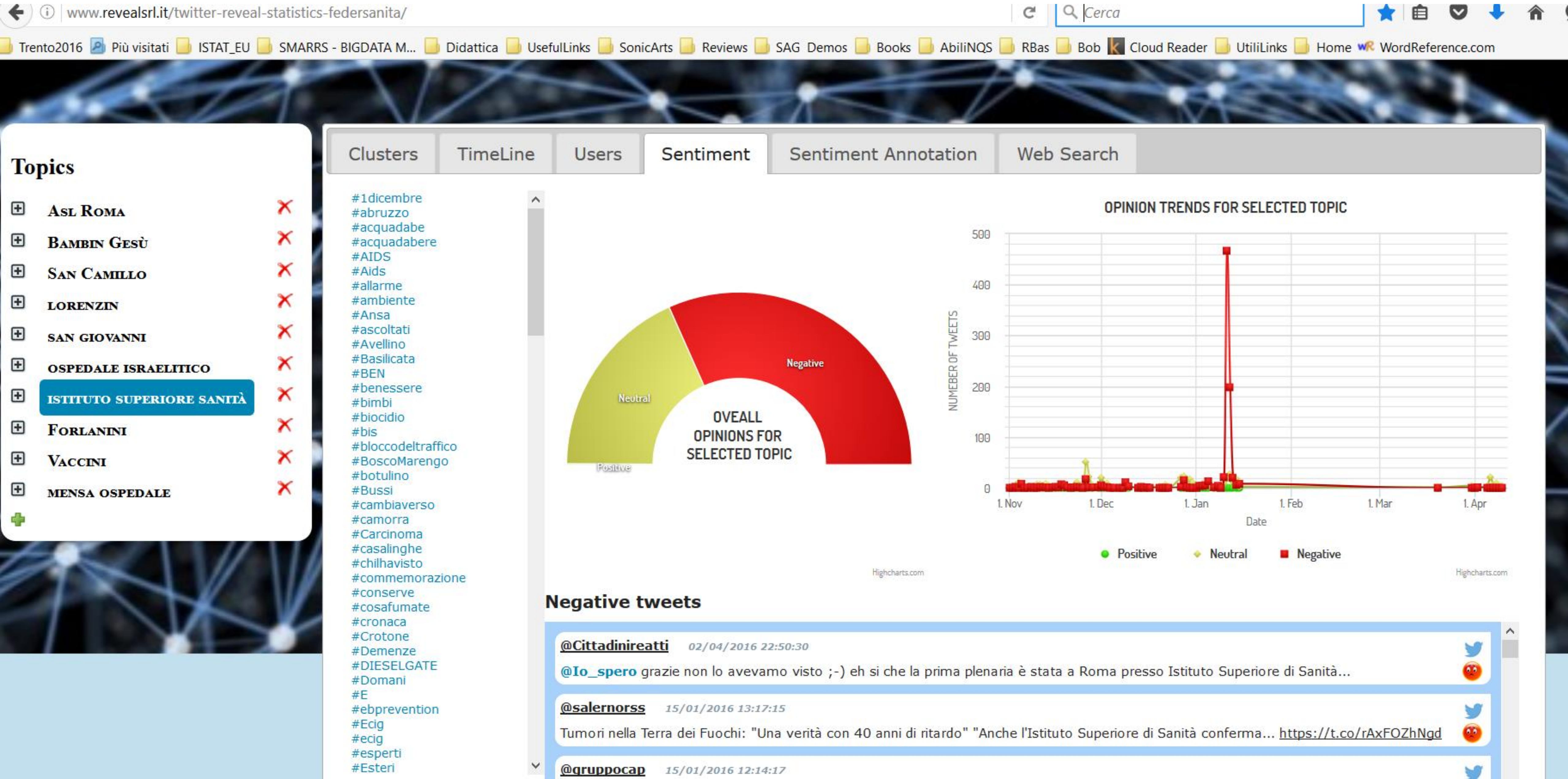
- Intelligenza Artificiale e Lingue parlate e scritte
  - Informazioni e Rappresentazioni coinvolte
  - Sfide (ri)correnti, battaglie (già) vinte e rischi inerenti …
- Elaborazione Automatica delle Lingue: Modelli, Metodi e *Risultati*
- *break*
- Ruolo delle Tecnologie dell’Apprendimento ed Applicazioni:
  - Sviluppo Automatico di Dizionari, Lessici Semantici ed Ontologie
  - Riconoscimento di fenomeni semantici
  - Trattamento Semantico della Documentazione Investigativa
  - Sistemi Web-based di Opinion Mining, Market Watch & Brand Reputation Management

The typical Semantic Search Architecture




www.revealsrl.it/twitter-reveal-statistics-federsanita/
Cerca
Trento2016
Più visitati
ISTAT_EU
SMARRS - BIGDATA M...
Didattica
UsefulLinks
SonicArts
Reviews
SAG Demos
Books
AbiliNQS
RBas
Bob
Cloud Reader
UtiliLinks
Home
WordReference.com
Topics
Asl Roma
Bambin Gesù
San Camillo
lorenzin
san giovanni
ospedale israelitico
istituto superiore sanità
Forlanini
Vaccini
mensa ospedale
Clusters
TimeLine
Users
Sentiment
Sentiment Annotation
Web Search
istituto, sanità, terra
sanità, tumore, ricovero
sanità, istituto, emergenza
sanità, terra, fuoco
terra, sanità, fuoco
Clusters
tumore, terra, istituto (127)
tumore, sanità, ricovero (59)
tumore, fuoco, L' (45)
terra, istituto, sanità (43)
terra, sanità, fuoco (39)
terra, fuoco, istituto (35)
terra, sanità, fuoco (31)
#ULTIMORA, terra, leggio (29)
sanità, istituto, formazione (28)
tumore, fuoco, terra (28)
sanità, terra, fuoco (25)
#vaccini, #Presadiretta, walter (24)
terra, sanità, eccesso (23)
sanità, dato, istituto (23)


File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto
La Repubblica.it - News in...
http://www.reve...s-federsanita/
www.revealsrl.it/twitter-reveal-statistics-federsanita/
Cerca
Trento2016 Più visitati ISTAT_EU SMARRS - BIGDATA M... Didattica UsefulLinks SonicArts Reviews SAG Demos Books AbiliNQS RBas Bob Cloud Reader UtiliLinks Home WordReference.com
Topics
Asl Roma
Bambin Gesù
San Camillo
lorenzin
san giovanni
ospedale israelitico
istituto superiore sanità
Forlanini
Vaccini
mensa ospedale
Clusters
TimeLine
Users
Sentiment
Sentiment Annotation
Web Search
any (932)
#ULTIMORA (72)
#Terradeifuochi (47)
#terradeifuochi (47)
#tumori (35)
#vaccini (27)
#Presadiretta (24)
#AIDS (20)
#TerradeiFuochi (20)
#UltimOra (17)
#chilhavisto (15)
#Sanit (13)
#lacitt (10)
#allarme (9)
#salute (7)
#bimbi (6)
#inquinamento (6)
#neonati (6)
#parto (6)
#saluteinrete (6)
#sanit (6)
#ambiente (5)
#informazione (5)
#ISS (5)
#Remuzzi (5)
35
30
25
20
15
10
5
0
2015/10/31 00:00 2015/11/18 00:00 2015/12/06 00:00 2015/12/24 00:00 2016/01/10 00:00 2016/01/28 00:00 2016/02/15 00:00 2016/03/04 00:00 2016/03/22 00:00
#Terradeifuochi #terradeifuochi #tumori #vaccini #Presadiretta #AIDS #TerradeiFuochi #UltimOra
#Sanit #lacitt #neonati #sanit #Remuzzi #ItaliaSicura #msdsalute #Napoli #servitumilitari
#INFLUENZA #acquadabe #BoscoMarengo #saluteinrete #parto #ISS


www.revealsrl.it/twitter-reveal-statistics-federsanita/
Cerca
Trento2016 Più visitati ISTAT_EU SMARRS - BIGDATA M... Didattica UsefulLinks SonicArts Reviews SAG Demos Books AbiliNQS RBas Bob Cloud Reader UtiliLinks Home WordReference.com
Topics
ASL ROMA
BAMBIN GESÙ
SAN CAMILLO
LORENZIN
SAN GIOVANNI
OSPEDALE ISRAELITICO
ISTITUTO SUPERIORE SANITÀ
FORLANINI
VACCINI
MENSA OSPEDALE
Clusters
TimeLine
Users
Sentiment
Sentiment Annotation
Web Search
#1dicembre
#abruzzo
#acquadabe
#acquadabere
#AIDS
#Aids
#allarme
#ambiente
#Ansa
#ascoltati
#Avellino
#Basilicata
#BEN
#benessere
#bimbi
#biocidio
#bis
#bloccodeltraffico
#BoscoMarengo
#botulino
#Bussi
#cambiaverso
#camorra
#Carcinoma
#casalinghe
#chilhavisto
#commemorazione
#conserve
#cosafumate
#cronaca
#Crotone
#Demenze
#DIESELGATE
#Domani
#E
#ebprevention
#Ecig
#ecig
#esperti
#Esteri
Negative
Neutral
Positive
OVEALL OPINIONS FOR SELECTED TOPIC
OPINION TRENDS FOR SELECTED TOPIC
500
400
300
200
100
0
NUMEBER OF TWEETS
1. Nov
1. Dec
1. Jan
1. Feb
1. Mar
1. Apr
Date
Positive
Neutral
Negative
Highcharts.com
Highcharts.com
Negative tweets
@Cittadinireatti 02/04/2016 22:50:30
@Io_spero grazie non lo avevamo visto ;-) eh si che la prima plenaria è stata a Roma presso Istituto Superiore di Sanità...
@salernorss 15/01/2016 13:17:15
Tumori nella Terra dei Fuochi: "Una verità con 40 anni di ritardo" "Anche l'Istituto Superiore di Sanità conferma... https://t.co/rAxFOZhNgd
@gruppocap 15/01/2016 12:14:17

Conclusioni

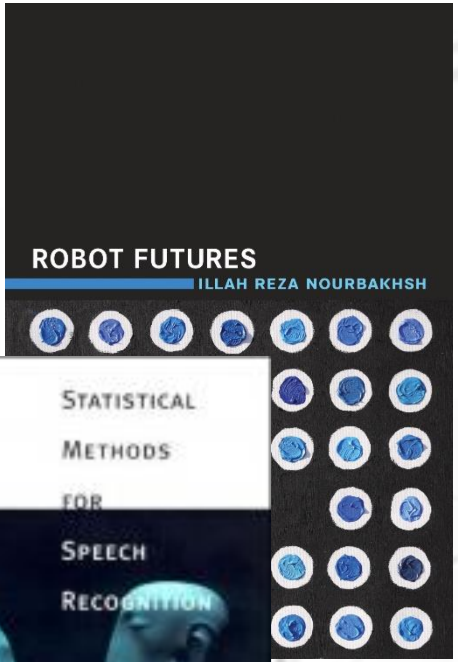
- I dati della odierna società della conoscenza sono opachi dal punto di vista epistemologico e l'intermediazione dei sistemi di calcolo deve sostenere processi complessi di interpretazione
- Le tecnologie del linguaggio e l'impulso loro dato dai metodi di Machine Learning possono svolgere un ruolo fondamentale nel sostenere in modo accurato i processi agenti sui Big Data e nel renderli economicamente sostenibili
- La tipica catena di elaborazione NLP è costituita da 4 fasi principali: Lexical Analysis, Syntactic Analysis, Semantic Analysis e Pragmatic (cioè Application-dependent) Analysis.
  - Le tecnologie di supporto alle tre fasi si basano su risorse (dizionari, lessici, grammatiche e basi di conoscenza) molto estese e dipendenti spesso dal dominio e dalla applicazione
  - Le tecnologie di Machine Learning consentono di abbattere i costi di messa a punto delle diverse componenti nei diversi domini di applicazione
  - Abbiamo approfondito alcuni compiti semantici (cioè legati alla fase di Analisi Semantica) come use cases nella aplicazione ML al NLP
    - Semantic Role Labeling
    - Named Entity Recognition and Classification

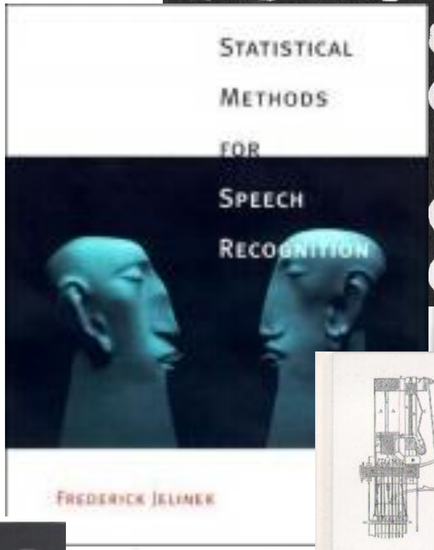
Conclusioni (2)

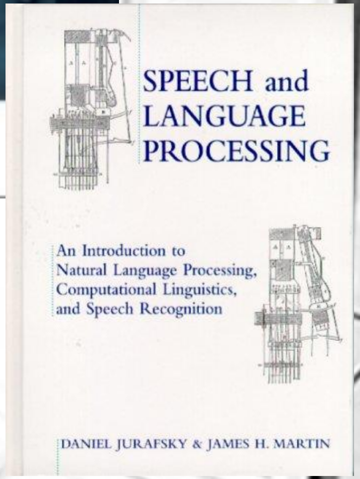
- I processi di AI (NLP&ML) costituiscono una branca attiva dell'Informatica che determina in modo rilevante il successo di processi innovativi della automazione in ambito industriale
  - Gestione Documentale
    - Metadatazione semantica
    - Indicizzazione
  - Semantic Search
    - Possibilità di gestire interrogazioni complesse (in NL) verso basi documentali estese indicizzate semanticamente in precedenza
  - Opinion Analysis & Brand Reputation
    - Analisi delle fonti aperte
    - Classificazione tematica ed emotiva
    - Business Intelligence sui livello dei contenuti e della emotività

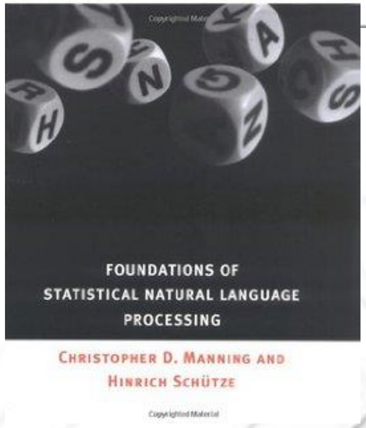
Riferimenti

- AI & Robotics. «Robot Futures», Ilah Reza Nourbakhsh, MIT Press, 2013
- NLP & ML:
  - «Statistical Methods for Speech Recognition», F. Jelinek, MIT Press, 1998
  - «Speech and Language Processing”, D. Jurafsky and J. H .Martin, Prentice-Hall, 2009.
  - “Foundations of Statistical Natural Language Processing, Manning & Schtze, MIT Press 2001.
- Sitografia:
  - SAG, Univ. Roma Tor Vergata: http://sag.art.uniroma2.it/
  - Reveal s.r.l.: http://www.revealsrl.it/









Riferimenti Bibliografici per il Corso

- Wordnet, Framenet:
  - Jurafsky-Martin «SPEECH and LANGUAGE PROCESSING» (http://www.cs.colorado.edu/~martin/slp.html), Chapt. 17, 19
- Semantic Role Labeling:
  - Jurafsky-Martin Book, Chapt. 10
  - Gildea, D. & Jurafsky, D. (2002). Automatic labeling of semantic roles. Computational Linguistics, 28:3, 245–288
  - S. Pradhan, et al., "Support Vector Learning for SRL", Journ. of Machine Learning, 60, 11–39, 2005, (URL: https://www.cs.colorado.edu/~martin/Papers/srl-ml-journal05.pdf).
  - Croce, D., Moschitti, A., & Basili, R. (2011). Structured lexical similarity via convolution Kernels on dependency trees. In EMNLP 2011: Conf. on Empirical Methods in NLP, 1034-1046, 2011.
- Named Entity Recognition
  - D.M. Bikel, R.L. Schwartz, R.M. Weischedel. 1999. An algorithm that learns what's in a name. Journ. of Machine Learning , 34(1-3):211–231